B65

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organizati n
International Bureau

(43) International Publication Date
28 December 2000 (28.12.2000)

PCT

(10) International Publication Number
WO 00/78921 A2

(51) International Patent Classification[7]: C12N

(21) International Application Number: PCT/US00/17322

(22) International Filing Date: 23 June 2000 (23.06.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/140,715 24 June 1999 (24.06.1999) US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 60/140,715 (CIP)
Filed on 24 June 1999 (24.06.1999)

(71) Applicants *(for all designated States except US)*: **THE JOHNS HOPKINS UNIVERSITY SCHOOL OF MEDICINE** [US/US]; Suite 906, 1111 Market Place, Baltimore, MD 21202 (US). **MASSACHUSETTS GENERAL HOSPITAL** [US/US]; Fruit Street, Boston, MA 02114 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **WORLEY, Paul** [US/US]; 17 Blythewood Road, Baltimore, MD 21210 (US). **TU, Jian, Cheng** [CN/US]; Apt. A, 6819 Blenheim Road, Baltimore, MD 21212 (US). **XIAO, Bo** [CN/US]; Apt. I, 3425 Plum Tree Drive, Ellicott City, MD 21042 (US). **SHENG, Morgan** [US/US]; Apt. 17D, 9 Hawthorne Place, Boston, MA 02114 (US).

(74) Agent: **HAILE, Lisa, A.**; Gray Cary Ware & Friedenrich LLP, Suite 1600, 4365 Executive Drive, San Diego, CA 92121-2189 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *Without international search report and to be republished upon receipt of that report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 00/78921 A2

(54) Title: SHANK PROTEINS AND METHODS OF USE THEREOF

(57) Abstract: Shank proteins and nucleotides encoding Shank proteins are provided. Also provided is a method for identifying a compound that modulates a cellular response mediated by Shank proteins. A method is further provided for identifying a compound that modulates cytoskeletal stability mediated by Shank proteins. A method is provided for identifying a compound that modulates synaptic receptor localization or synaptic receptor clustering, which are mediated by Shank proteins.

SHANK PROTEINS AND METHODS OF USE THEREOF

FIELD OF THE INVENTION

The present invention relates generally to protein-protein interactions and more specifically to molecules involved in mediating cytoskeletal stability and receptor localization.

BACKGROUND OF THE INVENTION

The mature central nervous system exhibits the capacity to alter cellular interactions as a function of the activity of specific neuronal circuits. This capacity is believed to underlie learning and memory storage, age-related memory loss, tolerance to and dependence on drugs of abuse, recovery from brain injury, epilepsy as well as aspects of postnatal development of the brain (Schatz, C., *Neuron*, 5:745, 1990). Currently, the role of activity-dependent synaptic plasticity is best understood in the context of learning and memory. Cellular mechanisms underlying activity-dependent plasticity are known to be initiated by rapid, transmitter-induced changes in membrane conductance properties and activation of intracellular signaling pathways (Bliss and Collingridge, *Nature*, 361:31, 1993). Several lines of evidence also indicate a role for rapid synthesis of mRNA and protein in long-term neuroplasticity.

Recent studies demonstrate that molecules that function together in cellular signaling networks are frequently clustered together in macromolecular complexes (see *e.g.*, Garner *et al.* (2000) Trends in Cell Biol. 10:274-280). For example, components of the MAP kinase pathway form a complex of cytosolic kinases with their specific substrates (Davis, *Mol. Reprod. Dev.* 42:459 (1995)). Similarly, proteins such as AKAP function as scaffolds for specific kinases and their substrates (Lester and Scott, *Recent Prog. Horm. Res.* 52:409 (1997)). Recently, a multi-PDZ containing protein was identified in *Drosophila* (termed InaD) that couples the membrane-associated, light-activated ion channel with its effector enzymes (Tsunoda *et al.*, *Nature* 388:243 (1997)). The biochemical consequence of this clustering is that the local concentrations of molecules that convey the signals between proteins are as high as possible. Consequently, signaling takes place efficiently. The clustering activity of these proteins is essential to normal function of the signaling cascade

(Lester and Scott, *supra* 1997; Tsunoda *et al.*, *supra* 1997). Accordingly, agents that alter these signaling complexes will modify the response due to transmitter or other form of cellular stimulation in a way that mimics more classical receptor agonists or antagonists.

NMDA (*N*-methyl-D-aspartate) receptors are a class of glutamate receptors that are highly permeable to calcium ions, and which activate a variety of signal transduction cascades. NMDA receptors are clustered at sites of excitatory synaptic contact between neurons in adult animals. Interactions between NMDA receptors and certain members of various families of intracellular proteins participate in localizing and concentrating receptors at excitatory synapses. The disruption or absence of appropriate excitatory synaptic transmission is implicated in a wide variety of diseases and disorders, in particular with respect to disorders of the central nervous system. Thus, the intracellular proteins that interact and maintain cytoskeletal stability, thereby maintaining NMDA receptors at excitatory synapses are important for normal functioning of the nervous system. Accordingly, there is a need in the art for intracellular proteins that mediate cytoskeletal stability and mediate receptor clustering.

SUMMARY OF THE INVENTION

The present invention provides a family of proteins that contain domains that can interact with other proteins. Through ankyrin domains, SH3 domains, PDZ domains, proline-rich domains and SAM domains, Shank proteins bind to GKAP proteins, PSD-95 proteins, cortactin and bind to other Shank proteins to form multimers. The present invention is based on the seminal discovery that Shank family proteins play a significant role in the post synaptic density cytoskeleton and that Shank family proteins regulate aspects of receptor clustering, in particular, clustering of NMDA receptors at synapses.

In one embodiment of the invention, there is provided a substantially pure polypeptide characterized as having an ankyrin domain, an SH3 domain; a PDZ domain; a proline-rich domain; and a SAM domain, and conservative variants thereof. The polypeptide can have an expression pattern in brain tissue. In addition, the

polypeptide interacts with intracellular proteins such as a cortactin protein, a PSD-95 protein, a Homer protein, a GKAP protein, and any combination thereof.

In another embodiment of the invention, there is provided a substantially pure polypeptide having an amino acid sequence as set forth in SEQ ID NO: 2, SEQ ID NO:4, SEQ ID NO:6, or conservative variants thereof. Also included are functional fragments thereof.

In yet another embodiment of the invention, there is provided an isolated polynucleotide selected from the group consisting of: (a) a polynucleotide encoding a polypeptide having an amino acid sequence as set forth in SEQ ID NO:2 or SEQ ID NO:4 or NO:6; a polynucleotide of (a), wherein T can be U; a polynucleotide complementary to (a) or (b); a polynucleotide having a nucleotide sequence as set forth in SEQ ID NO:1 or SEQ ID NO:3 or NO:5; degenerate variants of (a), (b), (c) or (d); a fragment of (a), (b), (c), (d) or (e) having at least 15 base pairs and that hybridizes to a polynucleotide encoding a polypeptide as set forth in SEQ ID NO:2, SEQ ID NO:4, SEQ ID NO:6; and a fragment of (a), (b), (c) (d) or (e) having at least 15 base pairs and that hybridizes to a polynucleotide encoding a polypeptide as set forth in amino acid residues 1 to 552 of SEQ ID NO:2 (Shank 1a) or residues 1 to 540 of SEQ ID NO:4 (Shank 3a).

In still a further embodiment of the invention, there is provided an isolated polynucleotide, wherein the nucleotide is at least 15 bases in length which hybridizes under moderately to highly stringent conditions to DNA encoding a polypeptide as set forth in SEQ ID NO:2, SEQ ID NO:4 or SEQ ID NO:6.

In an alternative embodiment of the invention, there is provided an antibody that binds to a Shank polypeptide or binds to an immunoreactive fragment thereof. The antibody can be polyclonal or monoclonal.

In yet another alternative embodiment of the present invention, there is provided an expression vector comprising a Shank polynucleotide, *e.g.*, SEQ ID NO:1 or SEQ ID NO:3, or complementary nucleotides thereof, and fragments thereof. The expression vector can be virus-derived or plasmid-derived.

In still a further embodiment of the invention, there is provided a method for producing a polypeptide by culturing a host cell containing a Shank polynucleotide

under conditions suitable for the expression of the polypeptide; and recovering the polypeptide from the host cell culture.

In another embodiment of the invention, there is provided a transgenic non-human animal having a transgene that expresses a Shank polypeptide chromosomally integrated into the germ cells of the animal.

In still another embodiment of the invention, there is provided a substantially pure polypeptide, wherein the polypeptide has a PDZ domain and interacts with amino acid sequence -X-T/S-R/K-L*, wherein X is any amino acid and L* is a carboxyl-terminal leucine residue. In a preferred embodiment, the polypeptide has the amino acid sequence -Q-T-R-L*.

In another embodiment of the invention, there is provided a computer readable medium having stored thereon a nucleic acid sequence selected from the group consisting of SEQ ID NO:1, SEQ ID NO:3, SEQ ID NO:5, and sequences substantially identical thereto, or a polypeptide sequence selected from the group consisting of SEQ ID NO:2, SEQ ID NO:4, SEQ ID NO:6 and sequences substantially identical thereto.

In another embodiment of the invention, there is provided a computer system comprising a processor and a data storage device wherein said data storage device has stored thereon a nucleic acid sequence selected from the group consisting of SEQ ID NO:1, SEQ ID NO:3, SEQ. ID NO:5 and sequences substantially identical thereto, or a polypeptide sequence selected from the group consisting SEQ ID NO:2, SEQ ID NO:4, SEQ ID NO:6 and sequences substantially identical thereto.

In yet another embodiment of the invention, there is provided a method for comparing a first sequence to a reference sequence wherein said first sequence is a nucleic acid sequence selected from the group consisting SEQ ID NO:1, SEQ ID NO:3, SEQ ID NO:5, and sequences substantially identical thereto, or a polypeptide sequence selected from the group consisting of SEQ ID NO:2, SEQ ID NO:4, SEQ ID NO:6 and sequences substantially identical thereto. The method comprises reading the first sequence and the reference sequence through use of a computer program which compares sequences, and determining differences between the first sequence and the reference sequence with the computer program.

In yet another embodiment of the invention there is provided a method for identifying a feature in a sequence wherein the sequence is selected from the group consisting of a nucleic acid sequence SEQ ID NO:1, SEQ ID NO:3, SEQ ID NO:5, sequences substantially identical thereto, or a polypeptide sequence SEQ ID NO:2, SEQ ID NO:4, SEQ ID NO:6 and sequences substantially identical thereto. The method includes reading the sequence through the use of a computer program which identifies features in sequences and identifying features in the sequences with the computer program.

In an additional embodiment of the invention, there is provided a method for identifying a compound that modulates a cellular response mediated by a Shank protein. The method includes incubating the compound and a cell expressing a Shank protein under conditions sufficient to permit the compound to interact with the cell, exposing the cell to conditions that activate the Shank protein and comparing a cellular response in the cell incubated with the compound with the cellular response of a cell not incubated with the compound wherein a difference in cellular response is indicative of a compound that modulates a cellular response mediated by a Shank protein.

In still another embodiment of the invention, there is provided a method for identifying a compound that modulates cytoskeletal stability. The method includes incubating the compound and a cell expressing a Shank protein under conditions sufficient to permit the compound to interact with the cell, exposing the cell to conditions sufficient to affect cytoskeletal stability, and comparing the cytoskeletal stability in the cell incubated with the compound with the cytoskeletal stability of a cell not incubated with the compound, thereby identifying a compound that modulates cytoskeletal stability.

In yet another embodiment of the invention, there is provided a method for identifying a compound that modulates receptor localization. The method includes incubating the compound and a cell expressing a Shank protein under conditions sufficient to permit the compound to interact with the cell, exposing the cell to conditions sufficient to affect receptor localization, and comparing the receptor localization in the cell incubated with the compound with the receptor localization of

a cell not incubated with the compound, thereby identifying a compound that modulates receptor localization.

In still a further embodiment of the invention, there is provided a method of identifying a compound that inhibits Shank protein activity. The method includes designing a potential inhibitor for Shank protein activity that will form non-covalent bonds with amino acids in a Shank protein binding site based upon the crystal structure co-ordinates of Shank protein binding domain and synthesizing the inhibitor. Then it can be determined whether the inhibitor inhibits Shank protein activity.

In yet another embodiment of the invention, there is provided a method for identifying a compound that affects the formation of cell surface receptors into clusters. The method includes incubating the compound and a cell expressing a Shank protein and a Homer protein under conditions sufficient to allow the compound to interact with the cell, determining the effect of the compound on the formation of cell-surface receptors into clusters, and comparing the formation of cell-surface receptors into clusters of the cell contacted with the compound with the formation of cell-surface receptors into clusters in a cell not contacted with the compound, thereby identifying a compound that affects the formation of cell-surface receptors into clusters.

In a further embodiment of the invention, there is provided a method of treating a disorder associated with glutamate receptors comprising administering to a subject in need thereof a therapeutically effective amount of a compound that modulates a Shank protein activity.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows the domain structure of Shank, and its interaction with GKAP1a. Figure 1A shows rat (r)and human (h) brain cDNA clones isolated from the yeast two-hybrid screen using GKAP1a C-terminal region (residues 591-666)as bait, aligned below a schematic of Shank protein (drawn to scale). Abbreviations are: Ank, Ankyrin repeats 1-7; SH3, Src homology 3 domain; PDZ, PSD-95/Dlg/ZO-1 domain; SAM, sterile alpha motif. Partial cDNAs from three related genes were isolated, termed Shank1, 2. and 3. Numbers in parentheses refer to the number of times each clone was isolated in the two-hybrid screen. Clone r11 contains an alternate

N-terminal sequence (hatched)preceding the PDZ domain. Figure 1B shows that the C terminus of GKAP1a interacts specifically with the PDZ domain of Shank. Interaction between GKAP1a (LexA fusions) and domains of Shank1 or PSD-95 (GAD fusions) were assayed by -gal/HIS3 induction in the yeast two-hybrid system. The C-terminal seven residues of GKAP1a (660-666) are sufficient to bind the Shank1 PDZ domain but show no interaction with the PDZ domains of PSD-95. Neither GKAP1b C-terminal splice variant (residues 602-627, terminating GQSK) nor the Kv1.4 C-terminal (residues 568-655. positive control for the PDZ domains of PSD-95) can bind the PDZ domain of Shank1 . Figure 1C shows the sequence requirements in the GKAP1a C terminus for interaction with Shank. The wild-type C-terminal sequence of GKAP1a (QTRL) is shown at top in bold. Single amino acid substitutions (bold, underlined) were introduced in the last four residues of GKAP1a (591-666). Interactions between mutant C termini and Shank1 (clone r8)were assayed described for Figure 1B.

Figure 2A shows the amino acid sequence alignment of Shank1a and Shank3a (SEQ ID NO:2 and 4, respectively). The sequence begins at the most likely translation start site based on Kozak consensus. Domains are underlined and labeled as in Figure 1A. Ankyrin repeats (r1-r7) are separated by black wedges. The Homer EVH-binding motif (see Tu et al., 1999) and cortactin SH3-binding motif are also underlined. Figure 2B shows the amino acid sequence of Shank2 (SEQ ID NO:6). Figures 2C, 2D and 2E show the nucleotide sequence of Shank1a, Shank2, and Shank3a, (SEQ ID NO:1, 5 and 3), respectively.

Figure 3 shows quantitative immunogold electron microscopic localization of Shank in the PSD. Figure 3A shows a quantitative analysis of the distribution of Shank immunogold particles at synapses. Fields containing Shank immunopositive synapses were digitized, and all gold particles within 150 nm of an active zone were analyzed. The distance of immunogold particles from the inner leaflet of the postsynaptic membrane in the axodendritic axis is plotted(in nm; 0 represents postsynaptic membrane). Figure 3B is a histogram showing the distance of gold particles from the center of the PSD in the lateral plane of the synapse (normalized by PSD length). Shank labeling peaks ~25 nm postsynaptic of the postsynaptic

membrane in the axodendritic axis and is evenly distributed in the lateral plane of the PSD

Figure 4 provides a diagram showing the regions of Shank3 used in certain GST pulldown assays (drawn to scale). The location of the putative cortactin binding sequence (KPPVPPKP)is indicated.

Figure 5 shows the importance of GKAP interaction on the synaptic localization of Shank, and glutamate-induced colocalization of cortactin and Shank in cultured neurons. Figure 5A shows the quantitation of immunocytochemical data obtained from neurons double-labeled for Shank and GKAP1a and GKAP1b, and GKAP. Neurons, transfected with GKAP1a and doubled labeled for GKAP) show high levels of GKAP staining on dendrites and punctate synaptic Shank staining, also along dendrites that is similar to untransfected neurons. GKAP1a-transfected neurons and GKAP1b-transfected neurons were also double-labeled for GKAP and PSD-95. The density of PSD-95 clusters is unaffected by overexpression of either GKAP1a or GKAP1b. The number of clusters of Shank and PSD-95 per 100 μm dendrite in neurons transfected with GKAP1a (gray bars)or GKAP1b (black), or in untransfected neurons (white) were counted using Metamorph software by a blind observer (see Example 8). Bars show mean ± SEM; * indicates $p < 0.01$ compared with either GKAP1a-transfected or untransfected neurons. In contrast, PSD-95 and synaptophysin clustering are not significantly different in GKAP1a versus GKAP1b-transfected neurons ($p = 0.52$ and $p = 0.35$, respectively). Figure 5B shows quantification of immunocytochemical data obtained from neurons double-labeled for Shank and cortactin, and with treatment by glutamate. Immunocytochemistry reveals colocalization of cortactin and Shank in growth of developing neurons. Mature neurons doubled labeled for cortactin and Shank reveal only a small fraction of puncta colocalized along dendrites. Most labeling does not overlap. After treatment with glutamate, (100 μm for 10 min), there is a marked increase in colocalization of Shank and cortactin such that most Shank immunoreactive puncta are also cortactin positive. Quantification of immunolabeling data shows the percent (pixel) area of cortactin labeling that overlaps with Shank labeling as determined using Metamorph colocalization software and plotted as mean ± SEM.

Figure 6 is a flow diagram illustrating a computer system, data retrieving device and display.

Figure 7 is a flow diagram illustrating one embodiment of process 200 for comparing a new nucleotide or protein sequence with a database of sequences in order to determine the homology levels between the new sequence and the sequences in the database.

Figure 8 is a flow diagram illustrating one embodiment of a process 250 in a computer for determining whether two sequences are homologous.

Figure 9 is a flow diagram illustrating one embodiment of a process 300 for comparing features in polynucleotide and polypeptide sequences.

DETAILED DESCRIPTION OF THE INVENTION

The identification of molecules regulating the aggregation of neurotransmitter receptors at synapses is central to understanding the mechanisms of neural development, synaptic plasticity and learning. The most well characterized model for the synaptic aggregation of ionotropic receptors is the neuromuscular junction. Early work showed that contact between the axon of a motor neuron and the surface of a myotube rapidly triggers the accumulation of preexisting surface acetylcholine receptors (Anderson and Cohen, *J Physiol.* 268:757-773, 1977; Frank and Fischbach, *J. Cell Biol.* 83:143-158, 1979). Subsequent work has shown that agrin, a complex glycoprotein secreted by the presynaptic terminal, activates a postsynaptic signal transduction cascade (reviewed by Colledge and Froehner, *Curr Opin Neurobiol* 8:357-63, 1998), that leads to receptor clustering by the membrane associated protein rapsyn.

Excitatory synaptic transmission in the mammalian brain is primarily mediated by the neurotransmitter glutamate acting on postsynaptic ionotropic glutamate receptors (particularly NMDA and AMPA receptors). In addition, glutamate stimulates a subset of metabotropic glutamate receptors (particularly the group I metabotropic glutamate receptors mGluR1a and mGluR5) concentrated in the postsynaptic membrane. The molecular mechanisms that underlie the postsynaptic localization and signaling capabilities of these glutamate receptors have been

intensely studied in recent years. An emerging theme is that the different classes of glutamate receptors (NMDA, AMPA, and group I metabotropic glutamate receptors) interact via their cytoplasmic tails with distinct intracellular anchoring/scaffold proteins (Sheng, 1997). The ionotropic receptors interact with specific PDZ domain proteins: NMDA receptors with the PSD-95/SAP90 family of proteins and AMPA receptors with GRIP/ABP/PICK1. On the other hand, mGluR1a and mGluR5 interact with the Homer/Vesl family of EVH domain proteins. These specific interactions may play a role in the synaptic targeting and cytoskeletal attachment of glutamate receptors. Perhaps more importantly, these anchoring proteins are thought to link their respective transmembrane receptors physically and functionally to the appropriate intracellular signaling pathways. For instance, PSD-95 may link NMDA receptors to neuronal nitric oxide synthase and a ras GTPase-activating protein (reviewed in Craven and Bredt, 1998), and Homer appears to couple mGluRs to the IP3 receptor (Tu et al., 1998). Despite recent advances, much remains to be learned about the molecular composition and the physiological functions of the protein complexes associated with PSD-95, GRIP, and Homer. Moreover, the apparent segregation of the different classes of glutamate receptors into parallel protein interaction pathways raises the question of whether the PSD-95-, GRIP-, and Homer-associated complexes cross-talk with each other via downstream protein interactions that have yet to be uncovered.

The postsynaptic density (PSD) can be visualized as an ultrastructural thickening of the postsynaptic membrane that is characteristic of excitatory synapses. Among the glutamate receptor complexes discussed above, the NMDA receptor/PSD-95 complex is the one most tightly associated with the PSD. In biochemical preparations of the PSD, NMDA receptors and PSD-95 are highly enriched and resistant to extraction by Triton X-100 and sarkosyl detergents, while AMPA receptors/GRIP and mGluRs/Homer are relatively soluble. It is possible that the components of the NMDA/receptor/PSD-95 complex comprise the major constituents of the core PSD remaining after extraction with strong detergents. Because they are likely to play critical roles in the structural organization of the synapse and in the transduction of NMDA receptor signals, these core PSD proteins are important to define and study.

Recently, a family of proteins (termed GKAP, SAPAP, or DAP) that is highly concentrated in the PSD and that binds to the guanylate kinase (GK) domain of PSD-95 has been identified. GKAP appears to be tightly associated with PSD-95; it can be immunoprecipitated from the brain in a complex with PSD-95 family proteins, and it is consistently colocalized with PSD-95 in neurons, even in the absence of associated NMDA receptors. The GKAP family of proteins contains at least four members and undergoes complex alternative splicing, but the physiological roles of these variants are unknown.

Shank is a family of proteins specifically enriched in the post-synaptic density (PSD) of excitatory synapses. Shank contains multiple protein interaction domains, including ankyrin repeats, the SH3 (Src homology 3) domain, the PDZ domain (*P*SD-95 *D*iscs large, *z*ona occludens 1 motif) the proline-rich domain, and the SAM (sterile alpha motif) domain. The PDZ domain of Shank mediates binding to the carboxyl terminus of GKAP (guanylate kinase-associated protein) and that this interaction is important in neurons for the synaptic localization of Shank. In addition, the SAM domain is responsible for multimerization of Shank, and the proline-rich region contains a specific binding site for cortactin, an actin cross-linking protein involved in regulation of the cortical actin cytoskeleton (Wu and Parsons, 1993; Huang et al., 1997). Shank also interacts specifically with Homer and group I metabotropic glutamate receptors (Tu et al., 1999). Thus, the Shank family proteins may be key organizers of the PSD, linking together the PSD- 95 and Homer-based complexes and allowing their interaction with modulators of the actin cytoskeleton.

Accordingly, in one embodiment of the invention, there is provided a substantially pure polypeptide characterized as having an ankyrin domain, an SH3 domain, a PDZ domain, a proline-rich domain and a SAM domain, and conservative variants thereof. The terms "conservative variation" and "substantially similar" as used herein denotes the replacement of an amino acid residue by another, biologically similar residue. Examples of conservative variations include the substitution of one hydrophobic residue such as isoleucine, valine, leucine or methionine for another, or the substitution of one polar residue for another, such as the substitution of arginine for lysine, glutamic acid for aspartic acid, or glutamine for asparagine, and the like. The terms "conservative variation" and "substantially similar" also include the use of

a substituted amino acid in place of an unsubstituted parent amino acid provided that antibodies raised to the substituted polypeptide also immunoreact with the unsubstituted polypeptide.

The Shank sequence has an abundance of motifs that are involved in binding to other proteins. These include ankyrin repeats, the SH3 domain, the PDZ domain, proline- rich motifs, and the SAM domain (see Figure 1A). Ankyrin repeats serve as protein interaction domains in proteins. In Shank proteins about seven ankyrin repeats in the ankyrin domain can be found. For example, in Shank1a (SEQ ID NO:2) the ankyrin domain containing seven ankyrin repeats is located between amino acid residues 104 and 1340. In Shank3a (SEQ ID NO:4), the ankyrin domain containing seven ankyrin repeats is located between amino acid residues 114 and 348.

Shank proteins contain another well-known protein-binding module, the SH3 domain. In Shank1a (SEQ ID NO:2), the SH3 domain includes amino acid residues 448 to 543 and in Shank3a (SEQ ID NO:4), residues 472-532.

Yet another domain found in Shank proteins is a PDZ domain which, as described herein (see Example 9), mediates Shank binding to the C terminus of GKAP. The PDZ domain of Shank has a distinctive binding specificity, preferring the hydrophobic residue leucine over valine at the very terminus of interacting proteins. This contrasts with the better known PDZ domains of PSD-95, which prefer valine at the 0 position. In addition, the Shank PDZ prefers positive charge over negative charge at the 21 position, whereas the best characterized ligands for the first two PDZ domains of PSD-95 (NMDA receptor NR2 subunits, and Shaker-type potassium channels) have a negatively charged aspartate in this position. A neutral amino acid may also be acceptable at 21 (see Tu et al., 1999 herein incorporated by reference). Based on sequence comparisons with PDZ domains of known binding specificity (Songyang et al., 1997) and the crystal structure of a PDZ–peptide complex, it is likely that the presence of isoleucine (residue 670 in Shank1a or 649 in Shank3a) at B8 might contribute to Shank's preference for leucine over valine at the 0 position. The negatively charged glutamate (residue 631 in Shank1a or 610 in Shank3a) at bC5 (instead of lysine in PDZ1/2 of PSD- 95) may contribute to Shank PDZ preference for a positive charge at the 21 position. The PDZ domain in

Shank1a, for example, is located at amino acid residues 587 to 684; in Shank3a, the PDZ domain is at residues 566 to 663.

Shank proteins are further characterized by having a proline rich domain. The proline-rich region contains a specific binding site for cortactin, an actin cross-linking protein involved in regulation of the cortical actin cytoskeleton.

The extensive region lying between the PDZ domain and the SAM domain of Shank is rich in proline (22% in Shank1, 16% in Shank3) and serine residues (16%in Shank1, 12% in Shank3). Proline-rich motifs often mediate protein-protein association, serving as binding sites for modules such as SH3, EVH, and WW domains (Bedford et al., 1997; Nguyen et al., 1998). As described herein, (see Example 4) there are at least two ligands for the proline-rich region of Shank: cortactin, which binds the -KPPVPPKP-motif with its SH3 domain, and Homer, which binds to the -PPXXF-motif (-PPLEF-in Shank1 (SEQ ID NO:2), -PPEEF-in Shank3(SEQ ID NO:3)), with its EVH domain.

Yet another domain that characterizes Shank proteins is the SAM domain which is found in a variety of signal transducing proteins, including Eph receptor tyrosine kinases. Interestingly, the SAM domain is found at the C terminus of all Eph receptors, the same position it occupies in the Shank polypeptides. Previous studies suggest that SAM domains can form homo-and/or hetero-oligomers. The crystal structure of the SAM domain from the EphB2 receptor has revealed two distinct interfaces for SAM-SAM interaction that would allow formation of an extended polymer of SAM domains. As described herein (see Example 12) full-length Shank can multimerize and the SAM domain of Shank is sufficient for self-association, suggesting that Shank can exist as an oligomer linked via its C-terminal SAM. In the context of the PSD, oligomerization of Shank SAM domains is significant for cross-linking multiple sets of protein complexes, such as the NMDA receptor/PSD-95/GKAP complex and the mGluR/Homer complex (Tu et al., 1999).

Shank proteins have an expression pattern in brain tissue. Immunofluorescence staining of cultured hippocampal neurons reveals a pattern of Shank immunoreactivity that is strikingly punctate and distributed along dendrites of neurons. The punctate Shank staining matched closely that of synaptophysin, GKAP,

PSD-95, and NRI, indicating the specific concentration of Shank in synapses. Shank-immunoreactive puncta show no colocalization with the GABAergic synaptic marker GAD indicating Shank's absence from inhibitory synapses. At the light microscopy level, therefore, Shank is a synaptic protein found specifically associated with excitatory synapses.

Shank proteins interact with certain intracellular proteins. Exemplary intracellular proteins that Shank proteins interact with includes, but is not limited to, a cortactin protein, a PSD-95 protein, a Homer protein, a GKAP protein.

Shank proteins interact directly with certain GKAP (guanylate kinase associated protein) proteins. As described herein (see Examples 6 and 7) Shank binds to GKAP1a, but not splice variant GKAP1b. Shank interacts with a PSD-95 protein; the interaction occurs indirectly, through certain GKAP proteins (see Examples 6 and 7).

Shank proteins also bind to cortactin (see Example 11). The interaction is mediated through the proline-rich domain of Shank proteins. Neurons, like other cells, undergo rearrangements of the cortical actin cytoskeleton in response to extracellular signals. The actin cytoskeleton of the dendritic spine is particularly dynamic and activity-dependent reorganization of the postsynaptic cytoskeleton may play a role in the plasticity of excitatory synapses. Little is understood, however, about the mechanisms that might couple synaptic stimulation to cytoskeletal changes in dendritic spines. As described herein, Shank binds to cortactin, a protein implicated in signaling to the actin cytoskeleton. Originally identified as a substrate of Src tyrosine kinase, cortactin is an F-actin-binding protein enriched in cell-matrix contact sites, membrane ruffles and lammelipodia of cultured cells, and in growth cones of neurons. The translocation of cortactin to the cell periphery is stimulated by the small GTPase Rac1, and its F-actin cross-linking activity is inhibited by Src tyrosine phosphorylation. Thus, a large body of evidence implicates cortactin in regulation of the actin cytoskeleton in dynamic regions of the cell periphery. These data indicate that cortactin can also play a role in neuronal synapses, based on the following findings: biochemically, cortactin is loosely associated with the PSD, and immunocytochemically, it colocalizes with Shank in a subset of synapses. Most

interestingly, there is a significant redistribution of cortactin to synaptic sites in response to glutamate stimulation. The glutamate-induced synaptic localization of cortactin is reminiscent of cortactin recruitment to the cortical cytoskeleton by growth factor stimulation of nonneural cells (Weed et al., 1998). Their coexistence in growth cones provides further evidence that Shank and cortactin can function at sites of active cytoskeletal remodeling in neurons. In mature synapses, a regulated Shank-cortactin interaction can be a mechanism for linking NMDA receptor activation to the control of the postsynaptic actin cytoskeleton.

Shank proteins can also interact with Homer proteins. Homer proteins, the products of neuronal immediate early genes, selectively bind the carboxy-termini of certain cell-surface receptors (*e.g.*, group 1 metabotropic receptors), certain intracellular receptors and binding proteins in addition to Shank proteins (*e.g.*, inositol trisphosphate receptors, ryanodine receptor, I42). Many forms of Homer proteins contain a "coiled-coil" structure in the carboxy-terminal domain which mediates homo- and heteromultimerization between Homer proteins. Homer plays a significant role in mediating receptor-activated calcium mobilization from internal stores and that Homer proteins regulate aspects of receptor clustering. Exemplary Homer proteins are Homer 1a, Homer 1b, Homer 1c, Homer 2a, Homer 2b and Homer 3 (see co-pending application PCT Application No. U.S. 99/18973, filed August 18, 1999, herein incorporated by reference in its entirety).

Shanks are highly related to CortBP1, a protein isolated by yeast two-hybrid screening with the SH3 domain of cortactin (Du et al., 1998). CortBP1 has been shown to colocalize with cortactin in membrane ruffles of cultured cells and in growth cones of cultured neurons (Du et al., 1998), analogous to our colocalization of Shank and cortactin in growth cones and synapses. Based on their similarity in primary structure and cell biological properties CortBP1 and Shanks as members of the same family of proteins.

In another embodiment of the invention, there is provided a substantially pure polypeptide, wherein the polypeptide has a PDZ domain and interacts with amino acid sequence -X-T/S-R/K-L*, wherein X is any amino acid and L* is a carboxyl-terminal leucine residue. In a preferred embodiment of the invention the amino acid sequence

is -Q-T-R-L*. The amino acid sequence -Q-T-R-L* is found, for example, in GKAP proteins.

Exemplary Shank polypeptides are set forth as SEQ ID NO:2 and SEQ ID NO:4 and conservative variants thereof.

Exemplary polynucleotides encoding Shank protein are set forth as SEQ ID NO: 1 and SEQ ID NO:3. The term "polynucleotide", "nucleic acid", "nucleic acid sequence", or "nucleic acid molecule" refers to a polymeric form of nucleotides at least 10 bases in length. By "isolated polynucleotide" is meant a polynucleotide that is not immediately contiguous with both of the coding sequences with which it is immediately contiguous (one on the 5' end and one on the 3' end) in the naturally occurring genome of the organism from which it is derived. The term therefore includes, for example, a recombinant DNA which is incorporated into a vector; into an autonomously replicating plasmid or virus; or into the genomic DNA of a prokaryote or eukaryote, or which exists as a separate molecule (*e.g.*, a cDNA) independent of other sequences. The nucleotides of the invention can be deoxyribonucleotides, ribonucleotides in which uracil (U) is present in place of thymine (T), or modified forms of either nucleotide. The nucleotides of the invention can be complementary to the deoxynucleotides or to the ribonucleotides. A polynucleotide encoding a Shank protein includes "degenerate variants", sequences that are degenerate as a result of the genetic code. There are 20 natural amino acids, most of which are specified by more than one codon. Therefore, all degenerate nucleotide sequences are included in the invention as long as the amino acid sequence of a polypeptide encoded by the nucleotide sequence of SEQ ID NO: 1 or SEQ ID NO:3 is functionally unchanged.

A nucleic acid molecule encoding a Shank protein includes sequences encoding functional Shank polypeptides as well as functional fragments thereof. As used herein, the term "functional polypeptide" refers to a polypeptide which possesses biological function or activity which is identified through a defined functional assay (*e.g.*, Examples 2, 4, 6, 7, 11 and 12), and which is associated with a particular biologic, morphologic, or phenotypic alteration in the cell. The term "functional fragments of Shank protein," refers to fragments of a Shank protein that retain a

Shank activity, *e.g.*, the ability to interact with intracellular proteins, and cell-surface receptors or mediate synaptic receptor localization or cytoskeletal stability, and the like. Additionally, functional Shank fragments may act as competitive inhibitors of Shank binding, for example, biologically functional fragments, for example, can vary in size from a polypeptide fragment as small as an epitope capable of binding an antibody molecule to a large polypeptide capable of participating in the characteristic induction or programming of phenotypic changes within a cell. Nucleotide fragments of the invention have at least 15 base pairs and hybridize to a polynucleotide encoding a polypeptide as set forth in SEQ ID NO:2 or SEQ ID NO:4.

An alternative embodiment provides nucleotide fragments having at least 15 base pairs and that hybridizes to a polynucleotide encoding a polypeptide as set forth in amino acid residues 1 to 552 of SEQ ID NO:2 or amino acid residues 1 to 540 of SEQ ID NO:4.

Yet another embodiment of the invention provides an isolated polynucleotide, wherein the nucleotide is at least 15 base pairs in length which hybridizes under moderately to highly stringent conditions to DNA encoding a polypeptide as set forth in SEQ ID NO:2 or SEQ ID NO:4. In nucleic acid hybridization reactions, the conditions used to achieve a particular level of stringency will vary, depending on the nature of the nucleic acids being hybridized. For example, the length, degree of complementarity, nucleotide sequence composition (*e.g.*, GC v. AT content), and nucleic acid type (*e.g.*, RNA v. DNA) of the hybridizing regions of the nucleic acids can be considered in selecting hybridization conditions. An additional consideration is whether one of the nucleic acids is immobilized, for example, on a filter.

An example of progressively higher stringency conditions is as follows: 2 x SSC/0.1% SDS at about room temperature (hybridization conditions); 0.2 x SSC/0.1% SDS at about room temperature (low stringency conditions); 0.2 x SSC/0.1% SDS at about 42 C (moderately stringent conditions); and 0.1 x SSC at about 68 C (highly stringent conditions). Washing can be carried out using only one of these conditions, *e.g.*, high stringency conditions, or each of the conditions can be used, *e.g.*, for 10-15 minutes each, in the order listed above, repeating any or all of the

steps listed. However, as mentioned above, optimal conditions will vary, depending on the particular hybridization reaction involved, and can be determined empirically.

Antibodies of the invention may bind to Shank provided by the invention to prevent normal interactions of Shank proteins. Binding of antibodies to Shank proteins can interfere with for example, cell-signaling, with receptor localization, with cytoskeletal stability, by interfering with intracellular protein binding. Binding of antibodies can interfere Shank protein binding to intracellular proteins, *e.g.*, to a cortactin protein, a PSD-95 protein, a Homer protein, a GKAP protein, and the like. Furthermore, binding to Shank proteins can interfere with cell-surface receptor clustering *e.g.* the clustering of NMDA receptors, mediated by Shank family proteins.

The antibodies of the invention can be used in any subject in which it is desirable to administer in vitro or in vivo immunodiagnosis or immunotherapy. The antibodies of the invention are suited for use, for example, in immunoassays in which they can be utilized in liquid phase or bound to a solid phase carrier. In addition, the antibodies in these immunoassays can be detectably labeled in various ways. Examples of types of immunoassays which can utilize antibodies of the invention are competitive and non-competitive immunoassays in either a direct or indirect format. Examples of such immunoassays are the radioimmunoassay (RIA) and the sandwich (immunometric) assay. Detection of the antigens using the antibodies of the invention can be done utilizing immunoassays which are run in either the forward, reverse, or simultaneous modes, including immunohistochemical assays on physiological samples. Those of skill in the art will know, or can readily discern, other immunoassay formats without undue experimentation.

The term "antibody" as used in this invention includes intact molecules as well as fragments thereof, such as Fab, F(ab')2, and Fv which are capable of binding to an epitopic determinant present in an invention polypeptide. Such antibody fragments retain some ability to selectively bind with its antigen or receptor.

Methods of making these fragments are known in the art. (See for example, Harlow and Lane, Antibodies: A Laboratory Manual, Cold Spring Harbor Laboratory, New York (1988), incorporated herein by reference). Monoclonal antibodies are made from antigen containing fragments of the protein by methods well known to

those skilled in the art (Kohler & Milstein, *Nature* 256:495 (1975); Coligan *et al.*, sections 2.5.1-2.6.7; and Harlow *et al.*, *Antibodies: A Laboratory Manual*, page 726 (Cold Spring Harbor Pub. 1988), which are hereby incorporated by reference. Briefly, monoclonal antibodies can be obtained by injecting mice with a composition comprising an antigen/ligand, verifying the presence of antibody production by analyzing a serum sample, removing the spleen to obtain B lymphocytes, fusing the B lymphocytes with myeloma cells to produce hybridomas, cloning the hybridomas, selecting positive clones that produce antibodies to the antigen, and isolating the antibodies from the hybridoma cultures. Monoclonal antibodies can be isolated and purified from hybridoma cultures by a variety of well-established techniques. Such isolation techniques include affinity chromatography with Protein-A Sepharose, size-exclusion chromatography, and ion-exchange chromatography. *See, e.g.*, Coligan *et al.*, sections 2.7.1-2.7.12 and sections 2.9.1-2.9.3; Barnes *et al.*, "Purification of Immunoglobulin G (IgG)" in *Methods In Molecular Biology*, VOL. 10, pages 79-104 (Humana Press 1992).

Antibodies which bind to an invention polypeptide can be prepared using an intact polypeptide or fragments containing small peptides of interest as the immunizing antigen. For example, it may be desirable to produce antibodies that specifically bind to the amino- or carboxyl-terminal domains of an invention polypeptide. For the preparation of polyclonal antibodies, the polypeptide or peptide used to immunize an animal is derived from translated cDNA or chemically synthesized and can be conjugated to a carrier protein, if desired. Commonly used carrier proteins which may be chemically coupled to the immunizing peptide include keyhole limpet hemocyanin (KLH), thyroglobulin, bovine serum albumin (BSA), tetanus toxoid, and the like.

Invention polyclonal or monoclonal antibodies can be further purified, for example, by binding to and elution from a matrix to which the polypeptide or a peptide to which the antibodies were raised is bound. Those of skill in the art will know of various techniques common in the immunology arts for purification and/or concentration of polyclonal antibodies, as well as monoclonal antibodies (See, for example, Coligan, *et al.*, Unit 9, Current Protocols in Immunology, Wiley Interscience, 1994, incorporated herein by reference).

The antibodies of the invention can be bound to many different carriers and used to detect the presence of an antigen comprising the polypeptides of the invention. Examples of well-known carriers include glass, polystyrene, polypropylene, polyethylene, dextran, nylon, amylases, natural and modified celluloses, polyacrylamides, agaroses and magnetite. The nature of the carrier can be either soluble or insoluble for purposes of the invention . Those skilled in the art will know of other suitable carriers for binding antibodies, or will be able to ascertain such, using routine experimentation.

There are many different labels and methods of labeling known to those of ordinary skill in the art. Examples of the types of labels which can be used in the present invention include enzymes, radioisotopes, fluorescent compounds, colloidal metals, chemiluminescent compounds, phosphorescent compounds, and bioluminescent compounds. Those of ordinary skill in the art will know of other suitable labels for binding to the antibody, or will be able to ascertain such, using routine experimentation.

Another technique which may also result in greater sensitivity consists of coupling the antibodies to low molecular weight haptens. These haptens can then be specifically detected by means of a second reaction. For example, it is common to use such haptens as biotin, which reacts with avidin, or dinitrophenyl, puridoxal, and fluorescein, which can react with specific antihapten antibodies.

In using the monoclonal and polyclonal antibodies of the invention for the in vivo detection of antigen, *e.g.*, a Shank protein, the detectably labeled antibody is given a dose which is diagnostically effective. The term "diagnostically effective" means that the amount of detectably labeled antibody is administered in sufficient quantity to enable detection of the site having the antigen comprising a polypeptide of the invention for which the antibodies are specific.

The concentration of detectably labeled antibody which is administered should be sufficient such that the binding to those cells having the polypeptide is detectable compared to the background. Further, it is desirable that the detectably labeled antibody be rapidly cleared from the circulatory system in order to give the best target-to-background signal ratio.

As a rule, the dosage of detectably labeled antibody for in vivo treatment or diagnosis will vary depending on such factors as age, sex, and extent of disease of the individual. Such dosages may vary, for example, depending on whether multiple injections are given, antigenic burden, and other factors known to those of skill in the art.

A polynucleotide agent can be contained in a vector, which can facilitate manipulation of the polynucleotide, including introduction of the polynucleotide into a target cell. The vector can be a cloning vector, which is useful for maintaining the polynucleotide, or can be an expression vector, which contains, in addition to the polynucleotide, regulatory elements useful for expressing the polynucleotide and, where the polynucleotide encodes a peptide, for expressing the encoded peptide in a particular cell. An expression vector can contain the expression elements necessary to achieve, for example, sustained transcription of the encoding polynucleotide, or the regulatory elements can be operatively linked to the polynucleotide prior to its being cloned into the vector.

An expression vector (or the polynucleotide) generally contains or encodes a promoter sequence, which can provide constitutive or, if desired, inducible or tissue specific or developmental stage specific expression of the encoding polynucleotide, a poly-A recognition sequence, and a ribosome recognition site or internal ribosome entry site, or other regulatory elements such as an enhancer, which can be tissue specific. The vector also can contain elements required for replication in a prokaryotic or eukaryotic host system or both, as desired. Such vectors, which include plasmid vectors and viral vectors such as bacteriophage, baculovirus, retrovirus, lentivirus, adenovirus, vaccinia virus, semliki forest virus and adeno-associated virus vectors, are well known and can be purchased from a commercial source (Promega, Madison WI; Stratagene, La Jolla CA; GIBCO/BRL, Gaithersburg MD) or can be constructed by one skilled in the art (see, for example, Meth. Enzymol., Vol. 185, Goeddel, ed. (Academic Press, Inc., 1990); Jolly, Canc. Gene Ther. 1:51-64, 1994; Flotte, J. Bioenerg. Biomemb. 25:37-42, 1993; Kirshenbaum et al., J. Clin. Invest. 92:381-387, 1993; each of which is incorporated herein by reference).

A polynucleotide useful in a method of the invention also can be operatively linked to tissue specific regulatory element, for example, a neuron specific regulatory element, such that expression of an encoded peptide agent is restricted to neurons in an individual, or to neurons in a mixed population of cells in culture, for example, an organ culture. For example, neuronal promoters such as the myelin basic protein promoter and other neuronal-specific promotes known to those of skill in the art may be used. Muscle-regulatory elements including, for example, the muscle creatine kinase promoter (Sternberg et al., Mol. Cell. Biol. 8:2896-2909, 1988, which is incorporated herein by reference) and the myosin light chain enhancer/promoter (Donoghue et al., Proc. Natl. Acad. Sci., *USA* 88:5847-5851, 1991, which is incorporated herein by reference) are well known in the art. A variety of other promoters have been identified which are suitable for up regulating expression in cardiac tissue. Included, for example, are the cardiac I-myosin heavy chain (AMHC) promoter and the cardiac I-actin promoter. Other examples of tissue-specific regulatory elements include, tissue-specific promoters, pancreatic (insulin or elastase), and actin promoter in smooth muscle cells. Through the use of promoters, such as milk-specific promoters, recombinant retroviruses may be isolated directly from the biological fluid of the progeny.

A Shank polynucleotide of the invention can be inserted into a vector, which can be a cloning vector or a recombinant expression vector. The term "expression vector" refers to a plasmid, virus or other vehicle known in the art that has been manipulated by insertion or incorporation of a polynucleotide, particularly, with respect to the present invention, a polynucleotide encoding all or a peptide portion of a shank protein. Such expression vectors contain a promoter sequence, which facilitates the efficient transcription of the inserted genetic sequence of the host. The expression vector generally contains an origin of replication, a promoter, as well as specific genes which allow phenotypic selection of the transformed cells. Vectors suitable for use in the present invention include, but are not limited to, the T7-based expression vector for expression in bacteria (Rosenberg, et al., Gene 56:125, 1987), the pMSXND expression vector for expression in mammalian cells (Lee and Nathans, J. Biol. Chem. 263:3521, 1988) and baculovirus-derived vectors for expression in insect cells. The DNA segment can be present in the vector operably linked to

regulatory elements, for example, a promoter, which can be a T7 promoter, metallothionein I promoter, polyhedrin promoter, or other promoter as desired, particularly tissue specific promoters or inducible promoters.

Viral expression vectors can be particularly useful for introducing a polynucleotide useful in a method of the invention into a cell, particularly a cell in a subject. Viral vectors provide the advantage that they can infect host cells with relatively high efficiency and can infect specific cell types. For example, a polynucleotide encoding a Shank protein or functional peptide portion thereof can be cloned into a baculovirus vector, which then can be used to infect an insect host cell, thereby providing a means to produce large amounts of the encoded protein or peptide portion. The viral vector also can be derived from a virus that infects cells of an organism of interest, for example, vertebrate host cells such as mammalian, avian or piscine host cells. Viral vectors can be particularly useful for introducing a polynucleotide useful in performing a method of the invention into a target cell. Viral vectors have been developed for use in particular host systems, particularly mammalian systems and include, for example, retroviral vectors, other lentivirus vectors such as those based on the human immunodeficiency virus (HIV), adenovirus vectors, adeno-associated virus vectors, herpesvirus vectors, vaccinia virus vectors, and the like (see Miller and Rosman, BioTechniques 7:980-990, 1992; Anderson et al., Nature 392:25-30 Suppl., 1998; Verma and Somia, Nature 389:239-242, 1997; Wilson, New Engl. J. Med. 334:1185-1187 (1996), each of which is incorporated herein by reference).

When retroviruses, for example, are used for gene transfer, replication competent retroviruses theoretically can develop due to recombination of retroviral vector and viral gene sequences in the packaging cell line utilized to produce the retroviral vector. Packaging cell lines in which the production of replication competent virus by recombination has been reduced or eliminated can be used to minimize the likelihood that a replication competent retrovirus will be produced. All retroviral vector supernatants used to infect cells are screened for replication competent virus by standard assays such as PCR and reverse transcriptase assays. Retroviral vectors allow for integration of a heterologous gene into a host cell

genome, which allows for the gene to be passed to daughter cells following cell division.

A polynucleotide, which can be contained in a vector, can be introduced into a cell by any of a variety of methods known in the art (Sambrook et al., Molecular Cloning: A laboratory manual (Cold Spring Harbor Laboratory Press 1989); Ausubel et al., Current Protocols in Molecular Biology, John Wiley and Sons, Baltimore, MD (1987, and supplements through 1995), each of which is incorporated herein by reference). Such methods include, for example, transfection, lipofection, microinjection, electroporation and, with viral vectors, infection; and can include the use of liposomes, microemulsions or the like, which can facilitate introduction of the polynucleotide into the cell and can protect the polynucleotide from degradation prior to its introduction into the cell. The selection of a particular method will depend, for example, on the cell into which the polynucleotide is to be introduced, as well as whether the cell is isolated in culture, or is in a tissue or organ in culture or *in situ*.

Introduction of a polynucleotide into a cell by infection with a viral vector is particularly advantageous in that it can efficiently introduce the nucleic acid molecule into a cell *ex vivo* or *in vivo* (see, for example, U.S. Patent No. 5,399,346, which is incorporated herein by reference). Moreover, viruses are very specialized and can be selected as vectors based on an ability to infect and propagate in one or a few specific cell types. Thus, their natural specificity can be used to target the nucleic acid molecule contained in the vector to specific cell types. As such, a vector based on an HIV can be used to infect T cells, a vector based on an adenovirus can be used, for example, to infect respiratory epithelial cells, a vector based on a herpesvirus can be used to infect neuronal cells, and the like. Other vectors, such as adeno-associated viruses can have greater host cell range and, therefore, can be used to infect various cell types, although viral or non-viral vectors also can be modified with specific receptors or ligands to alter target specificity through receptor mediated events.

A polynucleotide sequence encoding a Shank protein can be expressed in either prokaryotes or eukaryotes. Hosts can include microbial, yeast, insect and mammalian organisms. Methods of expressing polynucleotides having eukaryotic or viral sequences in prokaryotes are well known in the art, as are biologically functional

viral and plasmid DNA vectors capable of expression and replication in a host. Methods for constructing an expression vector containing a polynucleotide of the invention are well known, as are factors to be considered in selecting transcriptional or translational control signals, including, for example, whether the polynucleotide is to be expressed preferentially in a particular cell type or under particular conditions (see, for example, Sambrook et al., *supra*, 1989).

A variety of host cell/expression vector systems can be utilized to express a Shank receptor coding sequence, including, but not limited to, microorganisms such as bacteria transformed with recombinant bacteriophage DNA, plasmid DNA or cosmid DNA expression vectors; yeast cells transformed with recombinant yeast expression vectors; plant cell systems infected with recombinant virus expression vectors such as a cauliflower mosaic virus or tobacco mosaic virus, or transformed with recombinant plasmid expression vector such as a Ti plasmid; insect cells infected with recombinant virus expression vectors such as a baculovirus; animal cell systems infected with recombinant virus expression vectors such as a retrovirus, adenovirus or vaccinia virus vector; and transformed animal cell systems genetically engineered for stable expression. Where the expressed Shank protein is post-translationally modified, for example, by glycosylation, it can be particularly advantageous to select a host cell/expression vector system that can effect the desired modification, for example, a mammalian host cell/expression vector system.

Depending on the host cell/vector system utilized, any of a number of suitable transcription and translation elements, including constitutive and inducible promoters, transcription enhancer elements, transcription terminators, and the like can be used in the expression vector (Bitter et al., Meth. Enzymol. 153:516-544, 1987). For example, when cloning in bacterial systems, inducible promoters such as pL of bacteriophage λ, plac, ptrp, ptac (ptrp-lac hybrid promoter) and the like can be used. When cloning in mammalian cell systems, promoters derived from the genome of mammalian cells, for example, a human or mouse metallothionein promoter, or from mammalian viruses, for example, a retrovirus long terminal repeat, an adenovirus late promoter or a vaccinia virus 7.5K promoter, can be used. Promoters produced by recombinant DNA or synthetic techniques can also be used to provide for transcription of the inserted GDF receptors coding sequence.

In yeast cells, a number of vectors containing constitutive or inducible promoters can be used (see Ausubel et al., *supra*, 1987, see chapter 13; Grant et al., Meth. Enzymol. 153:516-544, 1987; Glover, DNA Cloning Vol. II (IRL Press, 1986), see chapter 3; Bitter, Meth. Enzymol. 152:673-684, 1987; see, also, The Molecular Biology of the Yeast *Saccharomyces* (Eds., Strathern et al., Cold Spring Harbor Laboratory Press, 1982), Vols. I and II). A constitutive yeast promoter such as ADH or LEU2 or an inducible promoter such as GAL can be used (Rothstein, DNA Cloning Vol. II (*supra*, 1986), chapter 3). Alternatively, vectors can be used which promote integration of foreign DNA sequences into the yeast chromosome.

Eukaryotic systems, particularly mammalian expression systems, allow for proper post-translational modifications of expressed mammalian proteins. Eukaryotic cells which possess the cellular machinery for proper processing of the primary transcript, glycosylation, phosphorylation, and advantageously, plasma membrane insertion of the gene product can be used as host cells for the expression of a Shank protein, or functional peptide portion thereof.

Mammalian cell systems which utilize recombinant viruses or viral elements to direct expression can be engineered. For example, when using adenovirus expression vectors, the Shank receptor coding sequence can be ligated to an adenovirus transcription/translation control complex, *e.g.*, the late promoter and tripartite leader sequence. Alternatively, the vaccinia virus 7.5K promoter can be used (Mackett et al., Proc. Natl. Acad. Sci., *USA* 79:7415-7419, 1982; Mackett et al., J. Virol. 49:857-864, 1984; Panicali et al., Proc. Natl. Acad. Sci., *USA* 79:4927-4931, 1982). Particularly useful are bovine papilloma virus vectors, which can replicate as extrachromosomal elements (Sarver et al., Mol. Cell. Biol. 1:486, 1981). Shortly after entry of this DNA into mouse cells, the plasmid replicates to about 100 to 200 copies per cell. Transcription of the inserted cDNA does not require integration of the plasmid into the host cell chromosome, thereby yielding a high level of expression. These vectors can be used for stable expression by including a selectable marker in the plasmid, such as, for example, the neo gene. Alternatively, the retroviral genome can be modified for use as a vector capable of introducing and directing the expression of the Shank protein gene in host cells (Cone and Mulligan, Proc. Natl. Acad. Sci., *USA* 81:6349-6353, 1984). High level expression can also be achieved

using inducible promoters, including, but not limited to, the metallothionein IIA promoter and heat shock promoters.

For long term, high yield production of recombinant proteins, stable expression is preferred. Rather than using expression vectors which contain viral origins of replication, host cells can be transformed with Shank protein cDNA controlled by appropriate expression control elements such as promoter, enhancer, sequences, transcription terminators, and polyadenylation sites, and a selectable marker. The selectable marker in the recombinant plasmid can confer resistance to the selection, and allows cells to stably integrate the plasmid into their chromosomes and grow to form foci, which, in turn can be cloned and expanded into cell lines. For example, following the introduction of foreign DNA, engineered cells can be allowed to grow for 1 to 2 days in an enriched media, and then are switched to a selective media. A number of selection systems can be used, including, but not limited to, the herpes simplex virus thymidine kinase (Wigler et al., Cell 11:223, 1977), hypoxanthine-guanine phosphoribosyltransferase (Szybalska and Szybalski, Proc. Natl. Acad. Sci., *USA* 48:2026, 1982), and adenine phosphoribosyltransferase (Lowy, et al., Cell 22:817, 1980) genes can be employed in tk^-, $hgprt^-$ or $aprt^-$ cells respectively. Also, antimetabolite resistance can be used as the basis of selection for dhfr, which confers resistance to methotrexate (Wigler, et al., Proc. Natl. Acad. Sci. *USA* 77:3567, 1980; O'Hare et al., Proc. Natl. Acad. Sci., *USA* 78: 1527, 1981); gpt, which confers resistance to mycophenolic acid (Mulligan and Berg, Proc. Natl. Acad. Sci., *USA* 78:2072, 1981); neo, which confers resistance to the aminoglycoside G-418 (Colberre-Garapin et al., J. Mol. Biol. 150:1, 1981); and hygro, which confers resistance to hygromycin (Santerre et al., Gene 30:147, 1984) genes. Additional selectable genes, including trpB, which allows cells to utilize indole in place of tryptophan; hisD, which allows cells to utilize histinol in place of histidine (Hartman and Mulligan, Proc. Natl. Acad. Sci., *USA* 85:8047, 1988); and ODC (ornithine decarboxylase) which confers resistance to the ornithine decarboxylase inhibitor, 2-(difluoromethyl)-DL-ornithine, DFMO (McConlogue, Curr. Comm. Mol. Biol. (Cold Spring Harbor Laboratory Press, 1987), also have been described.

When the host is a eukaryote, such methods of transfection of DNA as calcium phosphate coprecipitates, conventional mechanical procedures such as microinjection,

electroporation, insertion of a plasmid encased in liposomes, or virus vectors can be used. Eukaryotic cells can also be cotransformed with DNA sequences encoding the Shank proteins of the invention, and a second foreign DNA molecule encoding a selectable phenotype, such as the herpes simplex thymidine kinase gene. Another method is to use a eukaryotic viral vector, such as simian virus 40 (SV40) or bovine papilloma virus, to transiently infect or transform eukaryotic cells and express the protein. (Gluzman, Eukaryotic Viral Vectors (Cold Spring Harbor Laboratory Press, 1982)).

The invention provides a method for producing a polypeptide encoded by the nucleotide sequence of SEQ ID NO:1 or SEQ ID NO:3 or fragments thereof, including culturing the host cell under conditions suitable for the expression of the polypeptide and recovering the polypeptide from the host cell culture.

A Shank polypeptide or a fragment thereof, can be encoded by a recombinant or non-recombinant nucleic acid molecule and expressed in a cell. Preparation of a Shank polypeptide by recombinant methods provides several advantages. In particular, the nucleic acid sequence encoding the Shank polypeptide can include additional nucleotide sequences encoding, for example, peptides useful for recovering the Shank polypeptide from the host cell. A Shank polypeptide can be recovered using well known methods, including, for example, precipitation, gel filtration, ion exchange, reverse-phase, or affinity chromatography (see, for example, Deutscher et al., "Guide to Protein Purification" in Meth. Enzymol., Vol. 182, (Academic Press, 1990)). Such methods also can be used to purify a fragment of a Shank polypeptide, for example, a particular binding sequence, from a cell in which it is naturally expressed.

A recombinant nucleic acid molecule encoding a Shank polypeptide or a fragment thereof can include, for example, a protease site, which can facilitate cleavage of the Shank polypeptide from a non-Shank polypeptide sequence, for example, a tag peptide, secretory peptide, or the like. As such, the recombinant nucleic acid molecule also can encode a tag peptide such as a polyhistidine sequence, a FLAG peptide (Hopp et al., Biotechnology 6:1204 (1988)), a glutathione S-transferase polypeptide or the like, which can be bound by divalent metal ions, a specific antibody (U.S. Patent No. 5,011,912), or glutathione, respectively, thus facilitating recovery and purification of the

Shank polypeptide comprising the peptide tag. Such tag peptides also can facilitate identification of the Shank polypeptide through stages of synthesis, chemical or enzymatic modification, linkage, or the like. Methods for purifying polypeptides comprising such tags are well known in the art and the reagents for performing such methods are commercially available.

A nucleic acid molecule encoding a Shank polypeptide can be engineered to contain one or more restriction endonuclease recognition and cleavage sites, which can facilitate, for example, substitution of an element of the Shank polypeptide such as the selective recognition domain or, where present, a spacer element. As such, related Shank polypeptides can be prepared, each having a similar activity, but having specificity for different function-forming contexts. A restriction endonuclease site also can be engineered into (or out of) the sequence coding a peptide portion of the Shank polypeptide, and can, but need not change one or more amino acids encoded by the particular sequence. Such a site can provide a simple means to identify the nucleic acid sequence, based on cleavage (or lack of cleavage) following contact with the relevant restriction endonuclease, and, where introduction of the site changes an amino acid, can further provide advantages based on the substitution.

In another series of embodiments, the present invention provides transgenic animal models diseases or disorders associated with mutations in the Shank protein genes. The animal may be essentially any amphibian, reptile, fish, mammal, and the like. Preferably, the transgenic animal is mammalian including rats, mice, hamsters, guinea pigs, rabbits, dogs, cats, goats, sheep, pigs, and non-human primates. In addition, invertebrate models, including nematodes and insects, may be used for certain applications. The animal models are produced by standard transgenic methods including microinjection, transfection, or by other forms of transformation of embryonic stem cells, zygotes, gametes, and germ line cells with vectors including genomic or cDNA fragments, minigenes, homologous recombination vectors, viral insertion vectors and the like. Suitable vectors include vaccinia virus, adenovirus, adeno associated virus, retrovirus, liposome transport, neuraltropic viruses, Herpes simplex virus, and the like. The animal models may include transgenic sequences comprising or derived from Shank proteins including normal and mutant sequences,

intronic, exonic and untranslated sequences, and sequences encoding subsets of Shank proteins such as functional domains.

The major types of animal models provided include: (1) Animals in which a normal human Shank gene has been recombinantly introduced into the genome of the animal as an additional gene, under the regulation of either an exogenous or an endogenous promoter element, and as either a minigene or a large genomic fragment; in which a normal human Shank gene has been recombinantly substituted for one or both copies of the animal's homologous Shank gene by homologous recombination or gene targeting; and/or in which one or both copies of one of the animal's homologous Shank genes have been recombinantly "humanized" by the partial substitution of sequences encoding the human homologue by homologous recombination or gene targeting. (2) Animals in which a mutant human Shank gene has been recombinantly introduced into the genome of the animal as an additional gene, under the regulation of either an exogenous or an endogenous promoter element, and as either a minigene or a large genomic fragment; in which a mutant human Shank gene has been recombinantly substituted for one or both copies of the animal's homologous Shank gene by homologous recombination or gene targeting; and/or in which one or both copies of one of the animal's homologous Shank genes have been recombinantly "humanized" by the partial substitution of sequences encoding a mutant human homologue by homologous recombination or gene targeting. (3) Animals in which a mutant version of one of that animal's Shank genes has been recombinantly introduced into the genome of the animal as an additional gene, under the regulation of either an exogenous or an endogenous promoter element, and as either a minigene or a large genomic fragment; and/or in which a mutant version of one of that animal's Shank genes has been recombinantly substituted for one or both copies of the animal's homologous Shank gene by homologous recombination or gene targeting. (4) "Knock-out" animals in which one or both copies of one of the animal's Shank genes have been partially or completely deleted by homologous recombination or gene targeting, or have been inactivated by the insertion or substitution by homologous recombination or gene targeting of exogenous sequences.

In a preferred embodiment of the invention, there is provided a transgenic non-human animal having a transgene that expresses a Shank-encoding

polynucleotide chromosomally integrated into the germ cells of the animal. Animals are referred to as "transgenic" when such animal has had a heterologous DNA sequence, or one or more additional DNA sequences normally endogenous to the animal (collectively referred to herein as "transgenes") chromosomally integrated into the germ cells of the animal. The transgenic animal (including its progeny) will also have the transgene fortuitously integrated into the chromosomes of somatic cells.

Various methods to make the transgenic animals of the subject invention can be employed. Generally speaking, three such methods may be employed. In one such method, an embryo at the pronuclear stage (a "one cell embryo") is harvested from a female and the transgene is microinjected into the embryo, in which case the transgene will be chromosomally integrated into both the germ cells and somatic cells of the resulting mature animal. In another such method, embryonic stem cells are isolated and the transgene incorporated therein by electroporation, plasmid transfection or microinjection, followed by reintroduction of the stem cells into the embryo where they colonize and contribute to the germ line. Methods for microinjection of mammalian species is described in United States Patent No. 4,873,191. In yet another such method, embryonic cells are infected with a retrovirus containing the transgene whereby the germ cells of the embryo have the transgene chromosomally integrated therein. When the animals to be made transgenic are avian, because avian fertilized ova generally go through cell division for the first twenty hours in the oviduct, microinjection into the pronucleus of the fertilized egg is problematic due to the inaccessibility of the pronucleus. Therefore, of the methods to make transgenic animals described generally above, retrovirus infection is preferred for avian species, for example as described in U.S. Patent No. 5,162,215. If microinjection is to be used with avian species, however, a recently published procedure by Love *et al.*, (Biotechnology, 12, Jan 1994) can be utilized whereby the embryo is obtained from a sacrificed hen approximately two and one-half h after the laying of the previous laid egg, the transgene is microinjected into the cytoplasm of the germinal disc and the embryo is cultured in a host shell until maturity. When the animals to be made transgenic are bovine or porcine, microinjection can be hampered by the opacity of the ova thereby making the nuclei difficult to identify by traditional

differential interference-contrast microscopy. To overcome this problem, the ova can first be centrifuged to segregate the pronuclei for better visualization.

The non-human animals of the invention are murine typically (*e.g.*, mouse). The transgenic non-human animals of the invention are produced by introducing "transgenes" into the germline of the non-human animal. Embryonal target cells at various developmental stages can be used to introduce transgenes. Different methods are used depending on the stage of development of the embryonal target cell. The zygote is the best target for microinjection. The use of zygotes as a target for gene transfer has a major advantage in that in most cases the injected DNA will be incorporated into the host gene before the first cleavage (Brinster *et al., Proc. Natl. Acad. Sci. USA* 82:4438-4442, 1985). As a consequence, all cells of the transgenic non-human animal will carry the incorporated transgene. This will in general also be reflected in the efficient transmission of the transgene to offspring of the founder since 50% of the germ cells will harbor the transgene.

The term "transgenic" is used to describe an animal which includes exogenous genetic material within all of its cells. A "transgenic" animal can be produced by cross-breeding two chimeric animals which include exogenous genetic material within cells used in reproduction. Twenty-five percent of the resulting offspring will be transgenic *i.e.*, animals which include the exogenous genetic material within all of their cells in both alleles. 50% of the resulting animals will include the exogenous genetic material within one allele and 25% will include no exogenous genetic material.

In the microinjection method useful in the practice of the subject invention, the transgene is digested and purified free from any vector DNA *e.g.* by gel electrophoresis. It is preferred that the transgene include an operatively associated promoter which interacts with cellular proteins involved in transcription, ultimately resulting in constitutive expression. Promoters useful in this regard include those from cytomegalovirus (CMV), Moloney leukemia virus (MLV), and herpes virus, as well as those from the genes encoding metallothionine, skeletal actin, P-enolpyruvate carboxylase (PEPCK), phosphoglycerate (PGK), DHFR, and thymidine kinase. Promoters for viral long terminal repeats (LTRs) such as Rous Sarcoma Virus can also

be employed. Constructs useful in plasmid transfection of embryonic stem cells will employ additional regulatory elements well known in the art such as enhancer elements to stimulate transcription, splice acceptors, termination and polyadenylation signals, and ribosome binding sites to permit translation.

Retroviral infection can also be used to introduce transgene into a non-human animal, as described above. The developing non-human embryo can be cultured in vitro to the blastocyst stage. During this time, the blastomeres can be targets for retro viral infection (Jaenich, R., Proc. Natl. Acad. Sci USA 73:1260-1264, 1976). Efficient infection of the blastomeres is obtained by enzymatic treatment to remove the zona pellucida (Hogan, *et al.* (1986) in Manipulating the Mouse Embryo, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.). The viral vector system used to introduce the transgene is typically a replication-defective retro virus carrying the transgene (Jahner, *et al., Proc. Natl. Acad. Sci. USA* 82:6927-6931, 1985; Van der Putten, *et al., Proc. Natl. Acad. Sci USA* 82:6148-6152, 1985). Transfection is easily and efficiently obtained by culturing the blastomeres on a monolayer of virus-producing cells (Van der Putten, *supra*; Stewart, *et al., EMBO J.* 6:383-388, 1987). Alternatively, infection can be performed at a later stage. Virus or virus-producing cells can be injected into the blastocoel (D. Jahner *et al., Nature* 298:623-628, 1982). Most of the founders will be mosaic for the transgene since incorporation occurs only in a subset of the cells which formed the transgenic nonhuman animal. Further, the founder may contain various retro viral insertions of the transgene at different positions in the genome which generally will segregate in the offspring. In addition, it is also possible to introduce transgenes into the germ line, albeit with low efficiency, by intrauterine retroviral infection of the midgestation embryo (D. Jahner *et al., supra*).

A third type of target cell for transgene introduction is the embryonal stem cell (ES). ES cells are obtained from pre-implantation embryos cultured in vitro and fused with embryos (M. J. Evans *et al. Nature* 292:154-156, 1981; M.O. Bradley *et al., Nature* 309: 255-258, 1984; Gossler, *et al., Proc. Natl. Acad. Sci USA* 83: 9065-9069, 1986; and Robertson *et al., Nature* 322:445-448, 1986). Transgenes can be efficiently introduced into the ES cells by DNA transfection or by retro virus-mediated transduction. Such transformed ES cells can thereafter be combined with blastocysts

from a nonhuman animal. The ES cells thereafter colonize the embryo and contribute to the germ line of the resulting chimeric animal. (For review see Jaenisch, R., *Science* 240: 1468-1474, 1988).

"Transformed" means a cell into which (or into an ancestor of which) has been introduced, by means of recombinant nucleic acid techniques, a heterologous nucleic acid molecule. "Heterologous" refers to a nucleic acid sequence that either originates from another species or is modified from either its original form or the form primarily expressed in the cell.

"Transgene" means any piece of DNA which is inserted by artifice into a cell, and becomes part of the genome of the organism (*i.e.*, either stably integrated or as a stable extrachromosomal element) which develops from that cell. Such a transgene may include a gene which is partly or entirely heterologous (*i.e.*, foreign) to the transgenic organism, or may represent a gene homologous to an endogenous gene of the organism. Included within this definition is a transgene created by the providing of an RNA sequence which is transcribed into DNA and then incorporated into the genome. The transgenes of the invention include DNA sequences which encode Shank polypeptide-sense and antisense polynucleotides, which may be expressed in a transgenic non-human animal. The term "transgenic" as used herein additionally includes any organism whose genome has been altered by in vitro manipulation of the early embryo or fertilized egg or by any transgenic technology to induce a specific gene knockout. As used herein, the term "transgenic" includes any transgenic technology familiar to those in the art which can produce an organism carrying an introduced transgene or one in which an endogenous gene has been rendered non-functional or "knocked out".

Another embodiment of the invention provides a computer readable medium having store thereon a nucleic acid sequence selected from the group consisting of SEQ ID NO:1, SEQ ID NO:3, and sequences substantially identical thereto, or a polypeptide sequence selected from the group consisting of SEQ ID NO:2, SEQ ID NO:4, SEQ ID NO:6 and sequences substantially identical thereto.

A further embodiment of the invention provides a computer system comprising a processor and a data storage device wherein said date storage device has stored thereon a

nucleic acid sequence selected from the group consisting of SEQ ID NO:1, SEQ ID NO:3, and sequences substantially identical thereto, or a polypeptide sequence selected from the group consisting of SEQ ID NO:2, SEQ ID NO:4, and sequences substantially identical thereto. The computer system, additionally can contain a sequence comparison algorithm and a data storage device having at least one reference sequence stored on it. The sequence comparison algorithm comprises a computer program which indicates polymorphisms. The term "polymorphism", as used herein, refers to the existence of multiple alleles at a single locus. Polymorphism can be are several types including, for example, those that change DNA sequence but do not change protein sequence, those that change protein sequence without changing function, those that create proteins with a different activity, and those that create proteins that are non-functional.

Embodiments of the invention include systems (*e.g.*, internet based systems), particularly computer systems which store and manipulate the coordinate and sequence information described herein. One example of a computer system 100 is illustrated in block diagram form in Figure 4. As used herein, "a computer system" refers to the hardware components, software components, and data storage components used to analyze the coordinates and sequences as set forth in Table 1. The computer system 100 typically includes a processor for processing, accessing and manipulating the sequence data. The processor 105 can be any well-known type of central processing unit, such as, for example, the Pentium III from Intel Corporation, or similar processor from Sun, Motorola, Compaq, AMD or International Business Machines.

Typically the computer system 100 is a general purpose system that comprises the processor 105 and one or more internal data storage components 110 for storing data, and one or more data retrieving devices for retrieving the data stored on the data storage components. A skilled artisan can readily appreciate that any one of the currently available computer systems are suitable.

In one particular embodiment, the computer system 100 includes a processor 105 connected to a bus which is connected to a main memory 115 (preferably implemented as RAM) and one or more internal data storage devices 110, such as a hard drive and/or other computer readable media having data recorded thereon. In some embodiments, the computer system 100 further includes one or more data

retrieving device 118 for reading the data stored on the internal data storage devices 110.

The data retrieving device 118 may represent, for example, a floppy disk drive, a compact disk drive, a magnetic tape drive, or a modem capable of connection to a remote data storage system (*e.g.*, via the internet) etc. In some embodiments, the internal data storage device 110 is a removable computer readable medium such as a floppy disk, a compact disk, a magnetic tape, etc. containing control logic and/or data recorded thereon. The computer system 100 may advantageously include or be programmed by appropriate software for reading the control logic and/or the data from the data storage component once inserted in the data retrieving device.

The computer system 100 includes a display 120 which is used to display output to a computer user. It should also be noted that the computer system 100 can be linked to other computer systems 125a-c in a network or wide area network to provide centralized access to the computer system 100.

Figure 7 is a flow diagram illustrating one embodiment of a process 200 for comparing a new nucleotide or protein sequence with a database of sequences in order to determine the homology levels between the new sequence and the sequences in the database. The database of sequences can be a private database stored within the computer system 100, or a public database such as GENBANK that is available through the Internet.

The process 200 begins at a start state 201 and then moves to a state 202 wherein the new sequence to be compared is stored to a memory in a computer system 100. As discussed above, the memory could be any type of memory, including RAM or an internal storage device.

The process 200 then moves to a state 204 wherein a database of sequences is opened for analysis and comparison. The process 200 then moves to a state 206 wherein the first sequence stored in the database is read into a memory on the computer. A comparison is then performed at a state 210 to determine if the first sequence is the same as the second sequence. It is important to note that this step is not limited to performing an exact comparison between the new sequence and the first sequence in the database. Well-known methods are known to those of skill in the art

for comparing two nucleotide or protein sequences, even if they are not identical. For example, gaps can be introduced into one sequence in order to raise the homology level between the two tested sequences. The parameters that control whether gaps or other features are introduced into a sequence during comparison are normally entered by the user of the computer system.

Once a comparison of the two sequences has been performed at the state 210, a determination is made at a decision state 210 whether the two sequences are the same. Of course, the term "same" is not limited to sequences that are absolutely identical. Sequences that are within the homology parameters entered by the user will be marked as "same" in the process 200.

If a determination is made that the two sequences are the same, the process 200 moves to a state 214 wherein the name of the sequence from the database is displayed to the user. This state notifies the user that the sequence with the displayed name fulfills the homology constraints that were entered. Once the name of the stored sequence is displayed to the user, the process 200 moves to a decision state 218 wherein a determination is made whether more sequences exist in the database. If no more sequences exist in the database, then the process 200 terminates at an end state 220. However, if more sequences do exist in the database, then the process 200 moves to a state 224 wherein a pointer is moved to the next sequence in the database so that it can be compared to the new sequence. In this manner, the new sequence is aligned and compared with every sequence in the database.

It should be noted that if a determination had been made at the decision state 212 that the sequences were not homologous, then the process 200 would move immediately to the decision state 218 in order to determine if any other sequences were available in the database for comparison.

Figure 8 is a flow diagram illustrating one embodiment of a process 250 in a computer for determining whether two sequences are homologous. The process 250 begins at a start state 252 and then moves to a state 254 wherein a first sequence to be compared is stored to a memory. The second sequence to be compared is then stored to a memory at a state 256. The process 250 then moves to a state 260 wherein the first character in the first sequence is read and then to a state 262 wherein the first

character of the second sequence is read. It should be understood that if the sequence is a nucleotide sequence, then the character would normally be either A, T, C, G or U. If the sequence is a protein sequence, then it is preferably in the single letter amino acid code so that the first and sequence sequences can be easily compared.

A determination is then made at a decision state 264 whether the two characters are the same. If they are the same, then the process 250 moves to a state 268 wherein the next characters in the first and second sequences are read. A determination is then made whether the next characters are the same. If they are, then the process 250 continues this loop until two characters are not the same. If a determination is made that the next two characters are not the same, the process 250 moves to a decision state 274 to determine whether there are any more characters either sequence to read.

If there are not any more characters to read, then the process 250 moves to a state 276 wherein the level of homology between the first and second sequences is displayed to the user. The level of homology is determined by calculating the proportion of characters between the sequences that were the same out of the total number of sequences in the first sequence. Thus, if every character in a first 100 nucleotide sequence aligned with a every character in a second sequence, the homology level would be 100%.

Homology or identity is often measured using sequence analysis software (*e.g.*, Sequence Analysis Software Package of the Genetics Computer Group, University of Wisconsin Biotechnology Center, 1710 University Avenue, Madison, WI 53705). Such software matches similar sequences by assigning degrees of homology to various deletions, substitutions and other modifications. The terms "homology" and "identity" in the context of two or more nucleic acids or polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same when compared and aligned for maximum correspondence over a comparison window or designated region as measured using any number of sequence comparison algorithms or by manual alignment and visual inspection.

For sequence comparison, typically one sequence acts as a reference sequence, to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are entered into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. Default program parameters can be used, or alternative parameters can be designated. The sequence comparison algorithm then calculates the percent sequence identities for the test sequences relative to the reference sequence, based on the program parameters.

A "comparison window", as used herein, includes reference to a segment of any one of the number of contiguous positions selected from the group consisting of from 20 to 600, usually about 50 to about 200, more usually about 100 to about 150 in which a sequence may be compared to a reference sequence of the same number of contiguous positions after the two sequences are optimally aligned. Methods of alignment of sequence for comparison are well-known in the art. Optimal alignment of sequences for comparison can be conducted, *e.g.*, by the local homology algorithm of Smith & Waterman, Adv. Appl. Math. 2:482, 1981, by the homology alignment algorithm of Needleman & Wunsch, J. Mol. Biol. 48:443, 1970, by the search for similarity method of person & Lipman, Proc. Nat'l. Acad. Sci. USA 85:2444, 1988, by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by manual alignment and visual inspection. Other algorithms for determining homology or identity include, for example, in addition to a BLAST program (Basic Local Alignment Search Tool at the National Center for Biological Information), ALIGN, AMAS (Analysis of Multiply Aligned Sequences), AMPS (Protein Multiple Sequence Alignment), ASSET (Aligned Segment Statistical Evaluation Tool), BANDS, BESTSCOR, BIOSCAN (Biological Sequence Comparative Analysis Node), BLIMPS (BLocks IMProved Searcher), FASTA, Intervals & Points, BMB, CLUSTAL V, CLUSTAL W, CONSENSUS, LCONSENSUS, WCONSENSUS, Smith-Waterman algorithm, DARWIN, Las Vegas algorithm, FNAT (Forced Nucleotide Alignment Tool), Framealign, Framesearch, DYNAMIC, FILTER, FSAP (Fristensky Sequence Analysis Package), GAP (Global Alignment Program), GENAL, GIBBS, GenQuest, ISSC (Sensitive Sequence Comparison), LALIGN (Local Sequence Alignment), LCP (Local Content Program),

MACAW (Multiple Alignment Construction & Analysis Workbench), MAP (Multiple Alignment Program), MBLKP, MBLKN, PIMA (Pattern-Induced Multi-sequence Alignment), SAGA (Sequence Alignment by Genetic Algorithm) and WHAT-IF. Such alignment programs can also be used to screen genome databases to identify polynucleotide sequences having substantially identical sequences. A number of genome databases are available, for example, a substantial portion of the human genome is available as part of the Human Genome Sequencing Project (J. Roach, http://weber.u.Washington.edu/~roach/human_ genome_ progress 2.html) (Gibbs, 1995). At least twenty-one other genomes have already been sequenced, including, for example, *M. genitalium* (Fraser *et al.*, 1995), M. jannaschii (Bult *et al.*, 1996), *H. influenzae* (Fleischmann *et al.*, 1995), *E. coli* (Blattner *et al.*, 1997), and yeast (*S. cerevisiae*) (Mewes *et al.*, 1997), and *D. melanogaster* (Adams *et al.*, 2000). Significant progress has also been made in sequencing the genomes of model organism, such as mouse, *C. elegans*, and *Arabadopsis sp.* Several databases containing genomic information annotated with some functional information are maintained by different organization, and are accessible via the internet, for example, http://wwwtigr.org/tdb; http://www.genetics.wisc.edu; http://genome-www.stanford.edu/~ball; http://hiv-web.lanl.gov; http://www.ncbi.nlm.nih.gov; http://www.ebi.ac.uk; http://Pasteur.fr/other/biology; and http:// www.genome.wi.mit.edu.

One example of a useful algorithm is BLAST and BLAST 2.0 algorithms, which are described in Altschul *et al.*, Nuc. Acids Res. 25:3389-3402, 1977, and Altschul *et al.*, J. Mol. Biol. 215:403-410, 1990, respectively. Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al.*, *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always >0).

For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a word length (W) of 11, an expectation (E) of 10, M=5, N=-4 and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a word length of 3, and expectations (E) of 10, and the BLOSUM62 scoring matrix (see Henikoff & Henikoff, Proc. Natl. Acad. Sci. USA 89:10915, 1989) alignments (B) of 50, expectation (E) of 10, M=5, N= -4, and a comparison of both strands.

The BLAST algorithm also performs a statistical analysis of the similarity between two sequences (see, *e.g.*, Karlin & Altschul, Proc. Natl. Acad. Sci. USA 90:5873, 1993). One measure of similarity provided by BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a references sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.2, more preferably less than about 0.01, and most preferably less than about 0.001.

In one embodiment, protein and nucleic acid sequence homologies are evaluated using the Basic Local Alignment Search Tool ("BLAST") In particular, five specific BLAST programs are used to perform the following task:

(1) BLASTP and BLAST3 compare an amino acid query sequence against a protein sequence database;

(2) BLASTN compares a nucleotide query sequence against a nucleotide sequence database;

(3) BLASTX compares the six-frame conceptual translation products of a query nucleotide sequence (both strands) against a protein sequence database;

(4) TBLASTN compares a query protein sequence against a nucleotide sequence database translated in all six reading frames (both strands); and

(5) TBLASTX compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

The BLAST programs identify homologous sequences by identifying similar segments, which are referred to herein as "high-scoring segment pairs," between a query amino or nucleic acid sequence and a test sequence which is preferably obtained from a protein or nucleic acid sequence database. High-scoring segment pairs are preferably identified (*i.e.*, aligned) by means of a scoring matrix, many of which are known in the art. Preferably, the scoring matrix used is the BLOSUM62 matrix (Gonnet *et al.*, Science 256:1443-1445, 1992; Henikoff and Henikoff, Proteins 17:49-61, 1993). Less preferably, the PAM or PAM250 matrices may also be used (see, *e.g.*, Schwartz and Dayhoff, eds., 1978, *Matrices for Detecting Distance Relationships: Atlas of Protein Sequence and Structure*, Washington: National Biomedical Research Foundation). BLAST programs are accessible through the U.S. National Library of Medicine, *e.g.*, at www.ncbi.nlm.nih.gov.

The parameters used with the above algorithms may be adapted depending on the sequence length and degree of homology studied. In some embodiments, the parameters may be the default parameters used by the algorithms in the absence of instructions from the user.

A method is provided for identifying a compound that modulates a cellular response mediated by a Shank protein. The method includes incubating the compound and a cell expressing a Shank protein under conditions sufficient to permit the compound to interact with the cell. The effect of the compound on the cellular response is determined, either directly or indirectly, and a cellular response is then compared with a cellular response of a control cell. A suitable control includes, but is not limited to, a cellular response of a cell not contacted with the compound. The cell

may be any cell of interest, including but not limited to neuronal cells, glial cells, cardiac cells, bronchial cells, uterine cells, testicular cells, liver cells, renal cells, intestinal cells, cells from the thymus and spleen, placental cells, endothelial cells, endocrine cells including thyroid, parathyroid, pituitary and the like, smooth muscle cells and skeletal muscle cells. The term "incubating" includes conditions which allow contact between the test compound and the cell of interest. "Contacting" may include in solution or in solid phase.

The cellular response can be an increase in cytoskeletal stability or a decrease in cytoskeletal stability. Cytoskeletal stability can be assessed for example, by examining the formation and maintenance of intracellular protein interaction, cell-surface receptor clustering, and the like. Methods for demonstrating such cellular responses are well known in the art (*e.g.* biochemical methods and histological methods). (See Kornau *et al.* (1997) *Curr. Opin. Neurobiol.* 7:368-373; and Huganir *et al.* (2000) *Trends in Cell Biol.* 10:274-280, each of which are herein incorporated by reference in their entirety and Examples section for additional methodology).

Compounds which modulate a cellular response can include peptides, peptidomimetics, polypeptides, pharmaceuticals, chemical compounds and biological agents, for example. Antibodies, neurotropic agents, anti-epileptic compounds and combinatorial compound libraries can also be tested using the method of the invention. One class of organic molecules, preferably small organic compounds having a molecular weight of more than 50 and less than about 2,500 Daltons. Candidate agents comprise functional groups necessary for structural interaction with proteins, particularly hydrogen bonding, and typically include at least an amine, carbonyl, hydroxyl or carboxyl group, preferably at least two of the functional chemical groups. The candidate agents often comprise cyclical carbon or heterocyclic structures and/or aromatic or polyaromatic structures substituted with one or more of the above functional groups.

The test agent may also be a combinatorial library for screening a plurality of compounds. Compounds such as peptides identified in the method of the invention can be further cloned, sequenced, and the like, either in solution of after binding to a solid support, by any method usually applied to the isolation of a specific DNA

sequence Molecular techniques for DNA analysis (Landegren *et al.*, *Science* 242:229-237, 1988) and cloning have been reviewed (Sambrook *et al.*, Molecular Cloning: a Laboratory Manual, 2nd Ed.; Cold Spring Harbor Laboratory Press, Plainview, NY, 1998, herein incorporated by reference).

Candidate compounds are obtained from a wide variety of sources including libraries of synthetic or natural compounds. For example, numerous means are available for random and directed synthesis of a wide variety of organic compounds and biomolecules, including expression of randomized oligonucleotides and oligopeptides. Alternatively, libraries of natural compounds in the form of bacterial, fungal, plant and animal extracts are available or readily produced. Additionally, natural or synthetically produced libraries and compounds are readily modified through conventional chemical, physical and biochemical means, and may be used to produce combinatorial libraries. Known pharmacological agents may be subjected to directed or random chemical modifications, such as acylation, alkylation, esterification, amidification, etc., to produce structural analogs. Candidate agents are also found among biomolecules including, but not limited to: peptides, saccharides, fatty acids, steroids, purines, pyrimidines, derivatives, structural analogs or combinations thereof.

A variety of other agents may be included in the screening assay. These include agents like salts, neutral proteins, *e.g.*, albumin, detergents, *etc.* that are used to facilitate optimal protein-protein binding and/or reduce nonspecific or background interactions. Reagents that improve the efficiency of the assay, such as protease inhibitors, nuclease inhibitors, antimicrobial agents and the like may be used. The mixture of components are added in any order that provides for the requisite binding. Incubations are performed at any suitable temperature, typically between 4 and 40°C. Incubation periods are selected for optimum activity, but may also be optimized to facilitate rapid high-throughput screening. Typically between 0.1 and 10 h will be sufficient.

A method is provided for identifying a compound that modulates cytoskeletal stability by (a) incubating the compound and a cell expressing a Shank protein under conditions sufficient to permit the compound to interact with the cell; (b) exposing the

cell to conditions sufficient to affect cytoskeletal stability; and (c) comparing the cytoskeletal stability in the cell incubated with the compound with the cytoskeletal stability of a cell not incubated with the compound, thereby identifying a compound that modulates cytoskeletal stability.

A method is provided for identifying a compound that modulates receptor localization by: (a) incubating the compound and a cell expressing a Shank protein under conditions sufficient to permit the compound to interact with the cell; (b) exposing the cell to conditions sufficient to affect receptor localization; and (c) comparing the receptor localization in the cell incubated with the compound with the receptor localization of a cell not incubated with the compound, thereby identifying a compound that modulates receptor localization.

A compound can modulate receptor localization by either stimulating the formation of cell-surface receptors into clusters, resulting in an increase in receptor synaptic clustering, or by inhibiting the recruitment of cell-surface receptors into clusters, resulting in a decrease in receptor synaptic clustering. When the effect is "inhibition", cell-surface clustering is decreased as compared with the level in the absence of the test compound. When the effect is "stimulation", cell-surface clustering is increased as compared to a control in the absence of the test compound.

Receptors for a variety of chemical substances mediate communication between cells. Two general types of receptors are found in multicellular organisms: cell-surface receptors and intracellular, *e.g.*, nuclear receptors. Some hormones such as steroids and thyroxine pass through cell membranes to bind to intracellular receptors. Other hormones and neurotransmitters cannot pass through the cell membrane and these molecules bind to cell surface receptors.

A "cell-surface receptor" is a protein, usually having at least one binding domain on the outer surface of a cell where specific molecules may bind to, activate, or block the cell surface receptor. Cell surface receptors usually have at least one extracellular domain, a membrane spanning region ("transmembrane") and an intracellular domain. Activation of a cell-surface receptor can lead to changes in the levels of various molecules inside the cell. Several types of cell-surface receptors have been identified in a variety of cell types, including ligand-gated receptors,

ligand-gated channels, voltage-activated receptors, voltage-activated channels, ion channels and the like.

One class of cell-surface receptor is excitatory amino acid receptors (EAA receptors) which are the major class of excitatory neurotransmitter receptors in the central nervous system. "EAA receptors" are membrane spanning proteins that mediate the stimulatory actions of glutamate and possibly other endogenous acidic amino acids. EAA are crucial for fast excitatory neurotransmission and they have been implicated in a variety of diseases including Alzheimer's disease, stroke schizophrenia, head trauma and epilepsy. EAA have also been implicated in the process of aging In addition, EAA are integral to the processes of long-term potentiation, one of the synaptic mechanisms underlying learning and memory. There are three main subtypes of EAA receptors: (1) the metabotropic or trans ACPD receptors; (2) the ionotropic NMDA receptors; and (3) the non-NMDA receptors, which include the AMPA receptors and kainate receptors.

Ionotropic glutamate receptors are generally divided into two classes: the NMDA and non-NMDA receptors. Both classes of receptors are linked to integral cation channels and share some amino acid sequence homology. GluR1-4 are termed AMPA (α-amino -3-hydroxy-5-methylisoxazole-4-propionic acid) receptors because AMPA preferentially activates receptors composed of these subunits, while GluR5-7 are termed kainate receptors as these are preferentially sensitive to kainic acid. Thus, an "AMPA receptor" is a non-NMDA receptor that can be activated by AMPA. AMPA receptors include the GluR1-4 family, which form homo-oligomeric and hetero-oligomeric complexes which display different current-voltage relations and Ca^{2+} permeability. Polypeptides encoded by GluR1-4 nucleic acid sequences can form functional ligand-gated ion channels. An AMPA receptor includes a receptor having a GluR1, GluR2, GluR3 or GluR4 subunit. NMDA receptor subtypes include class NR2B and NR2D, for example.

Metabotropic glutamate receptors are divided into three groups based on amino acid sequence homology, transduction mechanism and binding selectivity: Group I, Group II and Group III. Each Group of receptors contains one or more types of receptors. For example, Group I includes metabotropic glutamate receptors 1 and 5

(mGluR1 and mGluR5), Group II includes metabotropic glutamate receptors 2 and 3 (mGluR2 and mGluR3) and Group III includes metabotropic glutamate receptors 4, 6, 7 and 8 (mGluR4, mGluR6, mGluR7 and mGluR8). Each mGluR type may be found in several subtypes. For example, subtypes of mGluR1 include mGluR1a, mGluR1b and mGluR1c.

Group I metabotropic glutamate receptors represent a family of seven membrane spanning proteins that couple to G-proteins and activate phospholipase C (Nakanishi, 1994). Members of the family include mGluR1 and mGluR5. Activation of these receptors results in the hydrolysis of membrane phosphatidylinositol bisphosphate to diacylglycerol, which activates protein kinase C. and inositol trisphosphate, which in turn activates the inositol trisphosphate receptor to release intracellular calcium. (Aramori and Nakanishi, 1992; Joly *et al.*, 1995 Kawabata *et al.*, 1998)

In another aspect of the invention, there is provided a method of identifying a compound that inhibits Shank protein activity by designing a potential inhibitor for Shank protein activity that will form non-covalent bonds with amino acids in a Shank protein binding site based upon the crystal structure co-ordinates of Shank protein binding domain; synthesizing the inhibitor; and determining whether the inhibitor inhibits Shank protein activity.

One aspect of the invention resides in the obtaining of crystals of a Shank protein of sufficient quality to determine the three dimensional (tertiary) structure of the protein by X-ray diffraction methods. The knowledge obtained concerning Shank proteins may be used in the determination of the three dimensional structure of the binding domain of Shank proteins. The binding domain can also be predicted by various computer models. Upon discovering the three-dimensional protein structure of the binding domain, small molecules which mimic the functional binding of Shank protein to its ligands can be designed and synthesized This is the method of "rational" drug design. Another approach to "rational" drug design is based on a lead compound that is discovered using high throughput screens; the lead compound is further modified based on a crystal structure of the binding regions of the molecule in question. Accordingly, another aspect of the invention is to provide material which is

a starting material in the rational design of drugs which mimic or prevents the action of Shank proteins.

The term "crystal structure coordinates" refers to mathematical coordinates derived from mathematical equations related to the patterns obtained on diffraction of a monochromatic beam of X-rays by the atoms (scattering centers) of a Shank protein molecule in crystal form. The diffraction data are used to calculate an electron density map of the repeating unit of the crystal. The electron density maps are used to establish the positions of the individual atoms within the unit cell of the crystal. The crystal structure coordinates of the Homer protein binding domain can be obtained using the electron density maps or by means of computational analysis.

The term "selenomethione substitution refers to the method of producing a chemically modified form of the crystal of Shank. The Shank protein is expressed by bacterial in media that is depleted in methionine and supplement in selenomethionine. Selenium is thereby incorporated into the crystal in place of methionine sulfurs.. The location(s) of selenium are determined by X-ray diffraction analysis of the crystal. This information is used to generate the phase information used to construct three-dimensional structure of the protein.

The term "heavy atom derivatization" refers to the method of producing a chemically modified form of the crystal of Shank. A crystal is soaked in a solution containing heavy metal atom salts or organometallic compounds, which can diffuse through the crystal and bind to the surface of the protein. The location(s) of the bound heavy metal atom(s) are determined by X-ray diffraction analysis of the soaked crystal. This information is used to generate the phase information used to construct three-dimensional structure of the protein.

Those of skill in the art understand that a set of structure coordinates determined by X-ray crystallography is not without standard error. The term "unit cell" refers to the basic paralleliped-shaped block. The entire volume of a crystal may be constructed by regular assembly of such blocks. The term "space group" refers to the arrangement of symmetry elements of a crystal.

The term "molecular replacement" refers to a method that involves generating a preliminary model of an Shank crystal whose structure coordinates are not known, by orienting and positioning a molecule whose structure coordinates are known. Phases are then calculated from this model and combined with observed amplitudes to give an approximate Fourier synthesis of the structure whose coordinates are known.

The crystal structure coordinates of Shank protein may be used to design compounds that bind to the protein and alter its physical or physiological properties in a variety of ways. The structure coordinates of the protein may also be used to computationally screen small molecule data bases for compounds that bind to the protein. The structure coordinates of Shank mutants (*e.g.*, missense mutations, deletion mutations, and the like, obtained by site-directed mutagenesis, by exposure to mutagenic agents, through selection of naturally occurring mutants, *etc.*) may also facilitate the identification of related proteins, thereby further leading to novel therapeutic modes for treating or preventing Shank-mediated conditions. A potential inhibitor is designed to form hydrogen bonds with certain amino acids of the Shank binding domain.

A method is further provided for treating a subject with a disorder associated with metabotropic or ionotropic glutamate receptors comprising administering to the subject a therapeutically effective amount of a compound that modulates Shank protein activity. In yet another embodiment, a method is provided for treating a subject with a disorder associated with Shank protein activity, comprising administering to the subject a therapeutically effective amount of a compound that modulates Shank protein activity.

Essentially, any disorder that is etiologically linked to a glutamate receptor or to a Shank protein could be considered susceptible to treatment with an agent that modulates Shank protein activity. The disorder may be a neuronal cell disorder. Examples of neuronal cell disorders include but are not limited to Alzheimer's disease, Parkinson's disease, stroke, epilepsy, neurodegenerative disease, Huntington's disease, and brain or spinal cord injury/damage, including ischemic injury. The disorder may also be a disorder of a cardiac disorder, a disorder of musculature, a renal disorder, a uterine disorder or a disorder of bronchial tissue. The

disorder may be epilepsy, glutamate toxicity, a disorder of memory, a disorder of learning or a disorder of brain development.

Detection of altered (decreased or increased) levels of "Shank protein activity" can be accomplished by hybridization of nucleic acids isolated from a cell of interest with a Shank polynucleotide of the invention. Analysis, such as Northern Blot analysis, are utilized to quantitate expression of Shank, such as to measure Shank transcripts. Other standard nucleic acid detection techniques will be known to those of skill in the art. Detection of altered levels of Shank can also accomplished using assays designed to detect Shank polypeptide. For example, antibodies or peptides that specifically bind a Shank polypeptide can be utilized. Analyses, such as radioimmune assay or immunohistochemistry, are then used to measure Shank, such as to measure protein concentration qualitatively or quantitatively.

Treatment can include modulation of Shank activity by administration of a therapeutically effective amount of a compound that modulates Shank or Shank protein activity. The term "modulate" envisions the suppression of Shank activity or expression when Shank is overexpressed or has an increased activity as compared to a control. The term "modulate" also includes the augmentation of the expression of Shank when it is underexpressed or has a decreased activity as compared to a control. The term "compound" as used herein describes any molecule, *e.g.*, protein, nucleic acid, or pharmaceutical, with the capability of altering the expression of Shank polynucleotide or activity of Shank polypeptide. Treatment may inhibit the interaction of a domain of Shank with its target protein, may increase the avidity of this interaction by means of allosteric effects, may block the binding activity of a domain of Shank or influence other functional properties of Shank proteins.

Candidate agents include nucleic acids encoding a Shank, or that interfere with expression of Shank, such as an antisense nucleic acid, ribozymes, and the like. Candidate agents also encompass numerous chemical classes wherein the agent modulates Shank expression or activity.

Where a disorder is associated with the increased expression of Shank, nucleic acid sequences that interfere with the expression of Shank can be used. In this manner, the coupling of cell-surface and intracellular receptors can be inhibited. This

approach also utilizes, for example, antisense nucleic acid, ribozymes, or triplex agents to block transcription or translation of Shank mRNA, either by masking that mRNA with an antisense nucleic acid or triplex agent, or by cleaving it with a ribozyme in disorders associated with increased Shank. Alternatively, a dominant negative form of Shank polypeptide could be administered.

When Shank is overexpressed, candidate agents include antisense nucleic acid sequences. Antisense nucleic acids are DNA or RNA molecules that are complementary to at least a portion of a specific mRNA molecule (Weintraub, 1990, *Scientific American*, **262**:40). In the cell, the antisense nucleic acids hybridize to the corresponding mRNA, forming a double-stranded molecule. The antisense nucleic acids interfere with the translation of the mRNA, since the cell will not translate a mRNA that is double-stranded. Antisense oligomers of about 15 nucleotides are preferred, since they are easily synthesized and are less likely to cause problems than larger molecules when introduced into the target cell. The use of antisense methods to inhibit the *in vitro* translation of genes is well known in the art (Marcus-Sakura, 1988, *Anal. Biochem.*, 172:289).

Use of an oligonucleotide to stall transcription is known as the triplex strategy since the oligomer winds around double-helical DNA, forming a three-strand helix. Therefore, these triplex compounds can be designed to recognize a unique site on a chosen gene (Maher, *et al.*, 1991, *Antisense Res. and Dev.*, 1(3):227; Helene, C., 1991, *Anticancer Drug Design*, 6(6):569).

Ribozymes are RNA molecules possessing the ability to specifically cleave other single-stranded RNA in a manner analogous to DNA restriction endonucleases. Through the modification of nucleotide sequences which encode these RNAs, it is possible to engineer molecules that recognize specific nucleotide sequences in an RNA molecule and cleave it (Cech, 1988, *J. Amer. Med. Assn.*, 260:3030). A major advantage of this approach is that, because they are sequence-specific, only mRNAs with particular sequences are inactivated.

There are two basic types of ribozymes namely, *tetrahymena*-type (Hasselhoff, 1988, *Nature*, **334**:585) and "hammerhead"-type. *Tetrahymena*-type ribozymes recognize sequences which are four bases in length, while

"hammerhead"-type ribozymes recognize base sequences 11-18 bases in length. The longer the recognition sequence, the greater the likelihood that the sequence will occur exclusively in the target mRNA species. Consequently, hammerhead-type ribozymes are preferable to *tetrahymena*-type ribozymes for inactivating a specific mRNA species and 18-based recognition sequences are preferable to shorter recognition sequences.

When a disorder is associated with the decreased expression of Shank, nucleic acid sequences that encode Shank can be used. An agent which modulates Shank expression includes a polynucleotide encoding a polypeptide of SEQ ID NO:2, SEQ ID NO:4, or a conservative variant thereof. Alternatively, an agent of use with the subject invention includes agents that increase the expression of a polynucleotide encoding Shank or an agent that increases the activity of Shank polypeptide.

The following examples are intended to illustrate but not limit the invention.

EXAMPLE1

Isolation and Primary Structure of the Shank Family of Proteins

Yeast Two-Hybrid Screen and Analysis of GKAP-Shank Interaction

Yeast two-hybrid screening and assays were performed as described in Bartel et al. (1993) and Niethammer and Sheng (1996) using the L40 yeast strain harboring HIS3 and -gal as reporter genes. GKAP1a residues 591-666 were subcloned into pBHA (LexA fusion vector) and used to screen ~1.5 x 10^6 clones from rat and human cDNA libraries constructed in pGADl0 (GAL4 activation domain vector, Clontech). For analysis of specificity and binding domains, desired cDNA segments were amplified by PCR with specific primers and subcloned into pBHA or pGAD10. Two-hybrid constructs of Kvl.4 and PSD-95 have been described (Kim et al., 1995). Shankl (SH3-PDZ) contains residues 469-691, and Shankl (PDZ) contains residues 684-691 of Shankla. Mutations of GKAP C-terminal residues were generated by PCR of GKAP1a (residues 591-666) using antisense primers containing specific nucleotide substitutions.

Shank Cloning and Plasmid Constructs Full-length Shankl and Shank3 cDNAs were obtained by standard hybridization screening of AZAP II rat cortical and hippocampal cDNA libraries (Stratagene) using digoxigenin-labeled DNA probes

from fragments initially isolated in the yeast two-hybrid screen. Full-length Shank1a cDNA constructs in GWI-CMV (British Biotechnology) were made by ligating restriction digested fragments from two-hybrid prey and λZAP II clones as follows: HindIII (nt 0, site introduced by short PCR)-BamHI (nt 1099) from clone 5-1; BamHI (nt 1099)-EcoRI (nt 2079) from clone r8/6; EcoRI (nt 2079)-BsawI (nt 3079) from prey clone r19; BsawI (nt 3079)-NotI (nt 3514) from r19/18; NotI (nt 3514)-AvrII (nt 4279)from clone 3-10; and AvrII (nt 4340)-SalI (nt 7090)from clone 1-3-1 (including 3'UTR). For experiments in COS7 cells, a Shank1a construct extending to clone 3-10 (diverging at nt 4759, residue 1509 of Shank1a, and thus omitting the SAM domain) was used because full-length Shank1a expressed poorly. For Shank3 expression constructs, see Tu et al. (Neuron (1999) 23:583-592) which is herein incorporated by reference in its entirety.

Isolation and Primary Structure of the Shank Family of Proteins To identify additional components in the postsynaptic PSD-95/GKAP complex, a search for GKAP-binding proteins by the yeast two-hybrid method was performed. Using as bait the C-terminal 76 residues of GKAP1a (originally termed GKAP in Kim et al., 1997) multiple copies of six distinct cDNAs from a screen of approximately 1.5×10^6 rat and human brain clones were obtained. No other interacting clones were isolated. Sequence analysis revealed that all six cDNAs were derived from three closely related genes. Four of the six interacting cDNA clones were overlapping sequences from the same gene; the remaining clones (r9 and h10) represented two distinct but highly homologous polypeptides (Figure 1A). The three GKAP-interacting genes are named Shank1-3, for the presence of an SH3 domain and multiple ankyrin repeats in the encoded polypeptides; the generic term Shank will be used for the whole family. The complete coding sequences of Shank1 and Shank3 were obtained following conventional hybridization screening of rat brain cDNA libraries (see *e.g.*, SEQ ID NO:1 and SEQ ID NO:3). During the cloning and sequencing of these cDNAs, complex alternative splicing of Shank1 (and other Shank family members) was revealed, with some variants resulting in severely truncated proteins (unpublished data). To prevent future confusion in nomenclature, the Shank1 and Shank3 splice variants presented in SEQ ID NO:1 and SEQ ID NO:3 are referred to as Shank1a and

Shank3a, respectively. Shank1a consists of 2088 residues, and Shank3a. 1740 residues.

Shank proteins share a common domain organization, consisting of seven ankyrin repeats near the N-terminus, followed by an SH3 domain, a PDZ domain, a long proline rich segment, and a SAM domain at the C-terminus (Figure 1A). All these motifs can be involved in protein-protein interactions, suggesting a scaffold function for the multidomain Shank polypeptides. Shank1a and Shank3a proteins (37% identical to each other over their entire length) are approximately 40% identical to the cortactin-binding protein CortBP1 (Du et al., 1998) over the region extending from the PDZ domain, where CortBP1 begins, to the C-terminus. A higher degree of identity is seen within the specified domains. CortBP1 does not contain the N-terminal ankyrin repeats or SH3 domains present in Shank1a and Shank3a, although it shares the PDZ, proline-rich. and SAM domains.

EXAMPLE 2

The C Terminus of GKAP Binds to the PDZ Domain of Shank

The PDZ domain is the only domain present in all interacting Shank clones isolated from the yeast two-hybrid screen (Figure 1A), suggesting that the Shank PDZ domain mediates interaction with the GKAP1a C-terminal bait construct. This was confirmed by deletion analysis in the yeast two-hybrid system; the isolated Shank1 PDZ domain was sufficient for binding to GKAP1a C-terminal residues 591-666 (Figure 1B). Furthermore, GKAP interaction is specific for the PDZ of Shank, since GKAP1a (residues 591-666)does not associate with the PDZ domains of PSD-95, Chapsyn-110/PSD-93, or CASK (Figure 1B; data not shown). Similarly, neither an N-terminal region of GKAP1a (residues 1-100) nor the Kv1.4 C-terminal tail could bind the Shank1 PDZ, even though these baits did interact, respectively, with the GK and PDZ domains of PSD-95 (Figure 1 B). The PDZ domain is also responsible for the interaction of Shank2 and Shank3 with GKAP1a (data not shown; see Tu et al., 1999).

PDZ domains typically bind to the last several residues at the C terminus of interacting proteins. Indeed, the last seven residues of GKAP1a (660-666; -PEAQTRL) interact with Shank1 as effectively and as specifically as the initial bait

GKAP1a (residues 591-666) (Figure 1 B) To define in detail the C-terminal residues involved in Shank PDZ binding, point mutants in the last four amino acids of GKAP1a were assayed for binding to Shank1 in the yeast two-hybrid system (Figure 1C). This analysis revealed that the final three residues of GKAP are important for specific binding to Shank. At the 0 position (the last amino acid), the wild-type residue leucine was preferred, but a conservative substitution with valine was tolerated; an alanine substitution at the very C terminus abolished interaction. At the -1 position, a positively charged amino acid (arginine or lysine) was greatly preferred over a negative charge (aspartate). At the -2 position, either threonine or serine supported Shank binding, but an alanine mutation abolished it. Changes at the -3 position (Q to E or Q to A) did not affect GKAP1a interaction with Shank (Figure 1C). This mutational analysis, which is not comprehensive, indicates that the C-terminal sequence preferred by Shank's PDZ domain is -X-T/S-R/K-L* (where X is any amino acid, and the asterisk represents a stop codon). Since all known members of the GKAP family (GKAPI SAPAP1-4) terminate with the same four amino acids (-QTRL) (Takeuchi et al., 1997), each member can presumably interact with Shank. However, it is noteworthy that splice variants of GKAP1 exist with alternative C termini, an example being GKAP1b (termed hGKAP in Kim et al., 1997). GKAP1b ends in the sequence GQSK* and does not interact with Shank (Figure 1 B).

EXAMPLE 3

Antibodies

Shank antibodies were raised in rabbits against a GST fusion of Shank1a residues 469-691 and affinity purified using a thioredoxin fusion of same region (these are termed Shank 56/8 antibodies). In some experiments, the IgG fraction purified by protein A-Sepharose was used (these are termed Shank 56/e antibodies). Identical bands were seen on rat brain immunoblots using both antibody preparations, and preincubation with a thioredoxin fusion of Shank1 (residues 469-591) abolished the signal of both antibodies (data not shown). In addition, an independent peptide antibody raised against a nonoverlapping region of Shank1 (residues 422-440) detected identical bands on rat brain immunoblots (unpublished data). The following antibodies have been described: anti-GKAP N1564 and C9589 (Kim et al., 1997; Naisbitt et al., 1997); guinea pig anti-PSD-95 (Kim et al., 1995): anti-cortactin mouse

monoclonal 4F11. a gift from Tom Parsons (Kanner et al., 1990). Anti-PSD-95 family mouse monoclonal antibodies (K28/86, Upstate Biotechnology) detect PSD-95, chapsyn-110/PSD-93. and SAP97 on immunoblots Other antibodies are described under the assays in which they are used.

EXAMPLE 4

Pulldown, Immunoprecipitation, and Biochemical Fractionation

For pulldown experiments, whole brain homogenate was extracted in 1% SDS and quenched in Tx-100 as described (Miller et al., 1996; Kim et al., 1997), except that quenching buffer contained (in mM): 130 KCl, 10 NaCl, 2 $MgCl_2$, 3 EGTA, 50 HEPES (pH 7.4) and 1% Tx-100. After 1 hr centrifugation at 100,000 x g, 120 µg extract was incubated with glutathione Sepharose 4B (Amersham Pharmacia Biotech) coupled to 6 µg GST or GST fusion (approximately 3 µl bed volume) for 2 hr at 4°C, followed by four washes in quenching buffer. Pulldowns and immunoprecipitations from transfected HEK293 cell extracts were performed as described in the accompanying paper (Tu et al.. 1999). For immunoprecipitation from COS7 cell extracts, cells (transfected as described herein) were washed and pelleted followed by resuspension in (in mM): 50 Tris (pH 7.4) 75 NaCl, 2.5 EGTA, 2.5 EDTA, 1% SDS, followed by l-fold dilution in 1%Tx-100. 50 Tris (pH 7.4) 150 NaCl, 2.5 EGTA, 2.5 EDTA plus protease inhibitors and 1 hr 16,000 x g centrifugation. Supernatants were incubated with 2 µg of control nonimmune IgG. Shank 56/e, or a 1:1 mixture of GKAP N1564 and C9589 antibodies. Extracts of forebrain synaptosomes for immunoprecipitation were prepared using either SDS (more efficient for GKAP coimmunoprecipitation with Shank) or DOC (more efficient for Shank coimmunoprecipitation with GKAP). Extraction of forebrain P2 in 1% DOC was performed as described (Dunah et al., 1998) followed by dialysis overnight into 0.1% Tx-1 00, 50 mM Tris (pH 7.4). Concurrently, 5 µg each antibody was preincubated overnight with 10 µl protein A-Sepharose. After clearing at 100,000 x g for 1 hr. dialyzed extract (50 µg protein) was incubated with antibody protein A in 100 µl 0.1%Tx-100, 50 mM Tris, (pH 7.4) for 2 hr at 4°C. Precipitates were washed four times with 1 ml of incubation buffer. For antigen competition controls, TRX-gk2.1 (see Kim et al., 1997) or TRX-GKAP1a (C-term) at 100 µg/ml concentration were present during all antibody incubation steps. For SDS extraction, the 1% Tx-100-

insoluble pellet from P2 was solubilized in 1% SDS plus 5 Mm ATP, 0.05% -mercaptoethanol, followed by 4-fold dilution in 1% Tx-100 quenching buffer plus 5 Mm ATP, 0.05% -mercaptoethanol. After centrifugation at 100,000 x g for 1 hr. soluble extract (300 µg protein) was incubated with 2.5 µg of control nonimmune IgG, Shank 56/e, or GKAP N1564 antibodies After 2 hr incubation at 4°C, 10 µl protein A-Sepharose beads was added for 2 hr. Pellets were washed four times in Tx-100 quenching buffer. Detergent extracted PSDI-III fractions were prepared as previously described (Cho et al., 1992). Immunoblotting was developed with enhanced chemiluminescence reagents (ECL. Amersham).

EXAMPLE 5

Transfection and Clustering in Heterologous Cells

COS-7 cells were transfected with Lipofectamine (GIBCO-BRL) on polylysine-coated coverslips (for clustering experiments) or in 100 mm tissue culture dishes (for immunoprecipitation experiments). Cells were fixed and permeabilized as described 24 hr after transfection (Kim et al., 1996; for concentrations of primary and secondary antibodies, see Example 8).

EXAMPLE 6

GKAP Binds Shank and Recruits Shank to PSD-95 Clusters in Heterologous Cells

To demonstrate biochemical association of GKAP and Shank proteins in a mammalian cell context, coimmunoprecipitation from transfected COS7 cells was performed (see Example 4). Since GKAP is known to associate with PSD-95 (Kim et al., 1997). PSD-95 was included in these experiments. In cells triply transfected with wild-type GKAP1a + Shank1 + PSD-95, both GKAP and Shank could be coimmunoprecipitated by antibodies specific for either protein, but not by control (nonimmune IgG) antibodies. GKAP antibodies also coprecipitated PSD-95, as expected. Antibodies to Shank brought down a significant amount of PSD-95 in addition to GKAP, implying the formation of a ternary complex containing Shank1GKAP/PSD-95. It was predicted that a C-terminal point mutation in GKAP1a (changing the last amino acid L666 to A) would abolish interaction with Shank but not interfere with binding to PSD-95, which is mediated by the N-terminal region of GKAP (Kim et al., 1997). In cells triply transfected with mutant GKAP1a (L666A) +

Shank1 + PSD-95, Shank was not coprecipitated by GKAP antibodies, even though the cognate antigen GKAP was efficiently brought down. Moreover, PSD-95 was not significantly coprecipitated with Shank antibodies, though its coimmunoprecipitation with GKAP remained robust. This experiment demonstrates that GKAP1a and Shank1 can associate in heterologous cells via a mechanism dependent on the C terminus of GKAP1a. Furthermore, GKAP1a can mediate the association of PSD-95 with Shank1. These findings are consistent with the formation of a ternary complex in which GKAP uses its N-terminal region to bind to PSD-95 and its C terminus to bind to Shank. Shank1 and GKAP1a also coimmunoprecipitate in the absence of PSD-95.

We have previously shown that PSD-95 and its relatives can cause the clustering of Shaker K^+ channels and NMDA receptors when coexpressed in COS7 cells (Kim et al., 1995, 1996) and that GKAP can be recruited into these clusters by binding to PSD-95 (Kim et al., 1997). To test whether GKAP might recruit Shank to PSD-95 clusters, coclustering in COS7 cells triply transfected with Shank1 + PSD-95 + GKAP1a or with Shank1 + PSD-95 +GKAP1b (the GKAP splice variant terminating in -GQSK* instead of -QTRL*) was examined. When expressed individually, Shank1, PSD-95, and GKAP1 are distributed typically in a diffuse cytoplasmic pattern in COS7 cells (data not shown). Shank1 and PSD-95 do not directly interact, and as expected, these two proteins do not cluster together when coexpressed (data not shown). In contrast, cells triply transfected with Shank1 + PSD-95 + GKAP1a formed plaque-like clusters in which Shank immunoreactivity colocalized precisely with PSD-95 and with GKAP (data not shown). These clusters have an appearance identical to PSD-95/Shaker K^+ channel coclusters or NMDA receptor/PSD-95 coclusters (Kim et al., 1995, 1996). In cells transfected with Shank1 + PSD-95 + GKAP1b, however, Shank immunoreactivity remains diffuse and does not colocalize with PSD-95. PSD-95 and GKAP form coaggregates in these cells, as they do even in the absence of Shank (data not shown). Thus, GKAP1a, but not GKAP1b, can mediate the coclustering of Shank1 with PSD-95 in heterologous cells. This is consistent with GKAP1a recruiting Shank1 via its specific C-terminal PDZ-binding sequence. Shank1 and GKAP1a do not form clusters in the absence of PSD-95 (data not shown). In summary, Shank1 can form coclusters with PSD-95

only in the presence of a GKAP splice variant that binds to Shank as well as to PSD-95. The GKAP1a-dependent coclustering of Shank1 and PSD-95 corroborates the coimmunoprecipitation results, indicating the formation in heterologous cells of a ternary complex of Shank/ GKAP/PSD-95 that is specified by the C-terminal sequence of GKAP.

EXAMPLE 7

Association of Shank with the GKAPIPSD-95 Complex In Vivo

GST pulldown experiments from rat brain extracts were performed to verify biochemical association of Shank and GKAP. Sepharose beads charged with GST fusion protein of Shank1(SH3-PDZ; residues 469-691) precipitated a large fraction of GKAP present in the offered extract. In addition to GKAP, GST-Shank1 (SH3-PDZ) brought down PSD-95, chapsyn-110/PSD-93, and NR1 (a subunit of NMDA receptors). These pulldowns were specific, since SAP97, synaptophysin, and glutamic acid decarboxylase (GAD) were not precipitated with GST-Shank1(SH3-PDZ), and GST alone pulled down none of these proteins.

In parallel, a GST fusion protein incorporating the C-terminal 76 amino acids of GKAP1a [GKAP1a(C-term)] specifically pulled down Shank proteins from brain extract. Multiple Shank bands are seen on immunoblots of the brain. The specificity of the Shank bands is supported by two pieces of evidence: first, these signals are abolished upon preincubation of antibodies with Shank immunogen, and second, two additional antibodies raised against independent regions of Shank yield essentially identical immunoblot patterns (unpublished data). The heterogeneity of Shank bands arises because the antibodies used in this study recognize multiple members of the Shank family and because each of the Shank genes undergoes complex alternative splicing (unpublished data). Although the GKAP1a(C-term) fusion protein pulled down Shank, it did not pull down GKAP. This is presumably because GST-GKAP1a(C-term) can only bind to Shank with an unoccupied PDZ domain and, hence, will not bind to Shank proteins already complexed with native GKAP in the brain extract. Consistent with such an interpretation, GST-GKAP1a(C-term) can bring down Homer protein, which binds to Shank at a site distinct from the PDZ domain (see Tu et al., 1999). In summary, these pulldown results confirm the specific binding between

GKAP and native Shank; furthermore, they indicate that Shank can bind to a native complex containing GKAP/PSD-95 and associated NMDA receptors.

To show that a ternary complex of Shank/GKAP/PSD-95 exists in vivo, coimmunoprecipitations from solubilized synaptosomal membranes were performed. Shank antibodies immunoprecipitated Shank proteins with high efficiency, and in addition they coprecipitated significant amounts of GKAP and PSD-95 and chapsyn-100/PSD-93. Since Shank does not interact directly with PSD-95 family proteins, this result is consistent with Shank existing in a ternary complex with GKAP/PSD95. Similarly, GKAP antibodies immunoprecipitated a large fraction of GKAP from the extracts, along with Shank, PSD-95, and chapsyn-110. Although this latter result does not prove the existence of the ternary complex, it does confirm the native association of GKAP with Shank and of GKAP with PSD-95 in the brain. None of the examined proteins was detectable in control IgG precipitates, indicating the specificity of the coimmunoprecipitations. Additionally, it is significant that SAP97 (a PSD-95 family protein; Muller et al., 1995) was not brought down with GKAP in either GST pulldown or antibody precipitation assays even though SAP97 has intrinsic binding affinity for GKAP (Kim et al., 1997). PSD-95 and chapsyn-110 cofractionate with GKAP as core components of the PSD, whereas SAP97 is segregated from GKAP in presynaptic and axonal compartments (Muller et al., 1995; Kim et al., 1996). The fact that GKAP and SAP97 do not coprecipitate thus offers reassurance that the detected protein-protein interactions are specific and not occurring artifactually after solubilization of the proteins. As a further test of specificity, the coimmunoprecipitation of Shank and GKAP can be blocked by competition with the specific antigen. GKAP antibodies raised against the N-terminal region of GKAP (termed N1564) immunoprecipitated GKAP and coprecipitated Shank and PSD-95 in the absence of any blocking antigens. The precipitation of GKAP and the coprecipitation of Shank and PSD-95 by N1564 antibodies were blocked by preincubation with the N-terminal GKAP fusion protein antigen but not by a fusion protein of the C-terminal region of GKAP.

EXAMPLE 8

Neur nal Culture Methodologies

Neuron Culture, Immunocytochemistry, and Immunoelectron Microscopy Low-density hippocampal neuronal cultures were prepared from E18 rats as described (Goslin and Banker, 1991). After 2-3 weeks in culture, neurons were fixed in methanol for 10 min at -20°C and incubated overnight at 4°C with primary antibodies at the following dilutions: Shank 56/e, 0.5 µg/ml; Synaptophysin SVP38 monoclonal (SIGMA), 1: 1000; GAD monoclonal (Boehringer Mannheim). 1 µg/ ml; PSD-95 guinea pig serum, 1:1000); GKAP N1564 or GKAP C9589, 1 µg/ml; NMDAR1 monoclonal 54.1 (PharMingen), 2.5 µg/ml. Cy3-and FITC-conjugated secondary antibodies (Jackson Immunoresearch) were used at dilutions of 1:500 and l: 100, respectively. Immunofluorescence was visualized with a Zeiss Axioskop microscope, and digital images were prepared for publication with Adobe Photoshop. Postembedding immunogold electron microscopy was performed and quantified as described (Phend et al., 1995; Naisbitt et al., 1997) using Shank antibody 56/e at 1 µg/ml. Only sparse scattered immunogold labeling was seen in the absence of primary antibody (data not shown).

Neuron Culture Transfection and Quantification of Immunolabeling For GKAP transfections, neuron cultures were prepared from trypsin dissociated E18-E19 hippocampi and plated on polylysine-coated coverslips in MEM containing 10% FCS. 25 µl/ml insulin, 100 µg/ml transferrin, 1 mM pyruvate, and 0.6%glucose. Transfections of GW1-GKAP1a or GW1-GKAP1b were performed at 3 days in vitro using the calcium phosphate method as described (Xia et al.. 1996). Neurons were fixed and processed for double-labeled immunostaining as described herein 7-10 days after transfection. GKAP1a-or GKAP1b-transfected neurons were easily recognized by their much higher levels of GKAP staining. GKAP1a-transfected neurons were double labeled for Shank (n =7) or PSD-95 (n =13) or synaptophysin (n =7) in addition to GKAP. Similarly, GKAP1b-transfected neurons were double-labeled for Shank (n =13) or PSD-95 (n =23) or synaptophysin (n =15). Images of transfected neurons and untransfected controls (n =7) were acquired using an interline coded CCD camera (Princeton Instruments) and analyzed by a blind observer using Metamorph software (Universal Imaging). For each neuron, immunolabelled puncta

having intensity above a blind user-defined threshold were counted by Metamorph software and normalized for dendritic length. An unpaired, two-tailed Student's t test on these sample arrays (having unequal variance by ANOVA) yielded $p < 0.01$ for Shank-labeled GKAP1b-transfected neurons when compared against either GKAP1 a-transfected or (untransfected cells $p=0.008$ or $p=0.006$, respectively). For cortactin redistribution experiments, low-density hippocampal neurons were prepared as described (Goslin and Banker, 1991). After 2 weeks in culture, glutamate (100 μM)was added directly to the medium for 10 min. Neurons were then fixed in methanol at 20°C and immunolabeled as described herein. Images were acquired as above, using constant camera exposure times within each experiment. Neurons were randomly chosen from treated and untreated cultures and subjected to a user-defined intensity threshold (kept constant within each experiment). Using this threshold, a colocalization index was obtained using Metamorph colocalization software, yielding the percent area (pixels) of cortactin signal that overlaps with Shank signal. Statistical significance was determined as above.

EXAMPLE 9

Biochemical and Electron Microscopic Evidence that Shank Is a Component of the PSD

To provide evidence that Shank is a component of the PSD in vivo, the fractionation of Shank proteins during biochemical purification of rat brain PSDs was examined. The heterogeneous set of Shank polypeptides was highly enriched in PSD preparations and was resistant to extraction by Triton X-100 and sarkosyl detergents. In fact, Shank enrichment during purification and detergent extraction of PSDs was similar to that of PSD-95 and GKAP. The finding that Shank is a core component of the PSD provides further evidence that Shank is an authentic component of the NMDA receptor/ PSD-95/GKAP complex.

Finally, postembedding immunoelectron microscopy was employed to investigate the subcellular location of Shank proteins in native brain tissue. Immunogold labeling for Shank in rat cerebral cortex is predominantly synaptic and associated with both axospinous and axodendritic asymmetric synapses. Most of the labeling is over the PSD, close to the postsynaptic membrane. Quantitative analysis confirmed that Shank is concentrated on the postsynaptic side of the synapse; the peak

of the distribution profile of Shank immunogold particles was approximately 25 nm inside the postsynaptic membrane (Figure 3A). Shank was relatively evenly distributed in the lateral plane of the synapse. These ultrastructural studies support the light microscopic and biochemical findings, confirming at high resolution that Shank is specifically concentrated in the PSD of excitatory synapses.

EXAMPLE 10

The GKAP-Shank Interaction Is Required for Shank Localization in Synapses

GKAP1b is a naturally occurring C-terminal splice variant of GKAP1 that binds to PSD-95 but not to Shank (see Example 4). Unlike GKAP1a, GKAP1b is unable to recruit Shank into PSD-95 clusters in heterologous cells. These two GKAP isoforms were exploited to explore the in vivo significance of the GKAP-Shank interaction. Overexpression of GKAP1a (the Shank-binding isoform) in cultured hippocampal neurons caused an increase in the density of Shank immunoreactive clusters (93 ± 16 clusters per 100 μm dendrite versus 69 ± 9 in untransfected neurons) that did not reach statistical significance ($p = 0.22$; Figure 4a). By contrast, overexpression of GKAP1b caused a marked and significant ($p < 0.01$) decrease in synaptic clusters of Shank (to only 31 ± 4 puncta per 100 μm dendrite)(Figure 4a). The number of synaptic PSD-95 clusters was not affected ($p = 0.52$) by either GKAP1a or GKAP1b overexpression (Figure 4a). Similarly, the density of synaptophysin puncta was not significantly altered (data not shown; $p = 0.35$). These findings indicate that the GKAP1a C terminus is functionally important in vivo for the targeting of Shank (but not of PSD-95) to synaptic sites. Together with the biochemical and immunostaining data, these dominant negative results support a physiological interaction between GKAP and Shank in neuronal synapses.

EXAMPLE 11

Shank Interaction with Cortactin

As noted herein, Shank1 and Shank3 show sequence similarity to the cortactin-binding protein CortBP1. CortBP1 contains a proline-rich motif (KPPVPPKP-) that mediates binding to the SH3 domain of cortactin (Du et al., 1998). An identical sequence in the proline-rich region of Shank3 (Figure 2, residues 1410-1417) is found that conforms to the cortactin SH3-binding consensus +PPψXKP determined by phage-displayed peptide library screening (+, ψ, and X signify basic,

aliphatic, or "any" residue, respectively; Sparks et al., 1996). Shank1 did not contain a sequence exactly matching this motif (the closest was -KPPLPPLP-, residues 1872-1879). To examine whether Shank3 can interact with cortactin, pulldown assays were performed with GST fusions of various Shank3 domains. Two constructs of Shank3 that contained the -KPPVPPKP-motif were able to bind cortactin expressed in HEK293 cells, while GST alone or fusions to other Shank domains could not. In the reverse direction, a GST fusion of full-length cortactin pulled down Shank3 (residues 1379-1740), while a GST fusion of cortactin with a specific SH3 domain deletion (cortactinΔSH3)was unable to do so. As further evidence for this interaction, full-length Shank3 was cotransfected with cortactin or cortactinΔSH3 into HEK cells. Antibodies to Shank were able to coimmunoprecipitate cortactin but not cortactinΔSH3. Pre-immune serum was unable to immunoprecipitate either protein. Thus, Shank3 can bind to cortactin in vitro and in heterologous cells, and the mode of binding is similar to the CortBP1-cortactin interaction (Du et al., 1998).

In the brain, cortactin is enriched in the PSD-I fraction (which has been extracted with Triton X-100), but not in PSD-III (extracted with Triton X-100 and sarkosyl). Therefore, cortactin may be weakly associated with the PSD, but unlike the Shank/GKAP/PSD-95 complex, cortactin does not behave as a core component of the PSD. Shank and cortactin do not coimmunoprecipitate from brain extracts, but this is perhaps not surprising given the differential detergent solubility of cortactin and Shank. Taken together, these biochemical results do not support a stable association of Shank and cortactin *in vivo*, but they are consistent with a regulated or low-affinity interaction between cortactin and Shank. To explore this further, the subcellular distribution of cortactin in cultured hippocampal neurons was examined. In developing neurons prior to synapse formation, cortactin and Shank are colocalized in growth cones of neuritic processes as has been shown previously for cortactin and CortBP1 (Du et al., 1998). In more mature neurons (2 weeks *in vitro*), the immunostaining pattern of cortactin was densely punctate but more widespread than that of Shank. Using a computer algorithm to quantitate the extent of area colocalization of immunofluorescent signals (see Example 10), it was found that a small fraction (6.3% ± 0.6%)of cortactin immunolabeling overlapped spatially with Shank (Figure 4B). Since Shank immunoreactivity is specifically clustered at

synapses, this minor punctate colocalization of cortactin and Shank likely occurs at synapses. Interestingly, after a 10 min stimulation of neurons with glutamate, cortactin redistributed to a more synaptic pattern, such that 25% ± 2.2% of cortactin immunolabeling colocalized with Shank ($p < 10^{-4}$) (Figure 4B). Since the algorithm used to determine this percentage does not take into account the intensity of immunostaining, and the brightest cortactin puncta colocalize with Shank, the actual mass of cortactin colocalizing with Shank is probably underestimated. The majority of Shank immunoreactive puncta colocalize with cortactin in glutamate-stimulated neurons. The colocalization data in primary neuron culture are consistent with an *in vivo* interaction of cortactin and Shank in growth cones and in a subset of synapses. Perhaps more interestingly, they suggest that cortactin may undergo an activity-dependent redistribution into synapses.

EXAMPLE 12

Multimerization of Shank

Several well-characterized scaffold proteins of the PSD show the capacity for homo-or heteromultimerization, including PSD-95 and chapsyn-110/PSD-93 (Kim et al., 1996; Hsueh et al., 1997), the Homer family of proteins (Xiao et al., 1998). and GRIP/ABP (Srivastava et al., 1998). To provide evidence for multimerization of Shank proteins, perhaps via the SAM domain-a domain known to mediate oligomerization (see Thanos et al., 1999, and references therein), the following studies were performed. To examine this issue, GST fusions of various regions of Shank3 were tested to determine which fusions could pull down a Shank3 fragment (residues 1379-1740) containing the SAM domain from extracts of transfected HEK293 cells. GST fusions of the C-terminal region of Shank3 (residues 1379-1740) or of the SAM domain alone (residues 1669-1740) were able to bind Shank3 (residues 1379-1740), while GST fusions of three other regions of Shank could not. Thus, regions of Shank3 containing the SAM domain are able to associate in vitro. In addition, when myc epitope-tagged full-length Shank3 was cotransfected with HA-tagged full-length Shank3 in HEK cells, anti-HA antibodies (but not nonimmune IgG) were able to coprecipitate mycShank3 with HA-Shank3 (Figure 7E). Anti-HA antibodies did not precipitate myc-Shank3 in the absence of HAShank3. Collectively,

these results imply that full-length Shank protein can multimerize and that the Shank SAM domain is sufficient to mediate this self-association.

References Cited

Allison, D. W.. Gelfand. V. I., Spector, I., and Craig, A. M. (1996). Role of actin in anchoring postsynaptic receptors in cultured hippocampal neurons: differential attachment of NMDA versus AMPA receptors. J. Neurosci. *16*, 2423-2436.

Bartel, P. L., Chien, C. T., Sternglanz, R.. and Fields. S. (1993). Using the Z-hybrid system to detect protein-protein interactions in cellular interactions in development: a practical approach, D. A. Hartley. ed. (Oxford: Oxford University Press).

Bedford, M. T.. Chart, D. C., and Leder. P. (1997). FBP WW domains and the Abl SH3 domain bind to a specific class of proline-rich ligands. EMBO J. *16*, 2376-2383.

Bork, P. (1993). Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile modules that cross phyla horizontally? Proteins *17*, 363-374.

Brakeman, P. R.. Lanahan. A. A., O'Brien. R..Roche. K., Barnes. C. A., Huganir. R. L., and Worley. P. F. (1997). Homer, a protein that selectively binds metabotropic glutamate receptors. Nature *386*, 284-266.

Cho. K.-O., Hunt, CA., and Kennedy, MB. (1992). The rat brain postsynaptic density fraction contains a homolog of the Drosophila discs-large tumor suppressor protein. Neuron *9*, 929-942.

Cohen, R. S., Blomberg, F., Berzins, K., and Siekevitx. P. (1977). The structure of postsynaptic densities isolated from dog cerebral cortex. I. Overall morphology and protein composition. J. Cell Biol. *74*, 161-203.

Cotman. C. W.. Banker, G.. Churchill, L.. and Taylor. D. (1974). Isolation of postsynaptic densities from rat brain. J. Cell Biol. *63*, 441-455.

Craven, S.. and Bredt. D. (1996). PDZ proteins organize synaptic signaling pathways. Cell *15*, 495-496.

Dong, H., O'Brien. R. J., Fung, ET., Lanahan, A. A., Worley, P. F., and Huganir, R. L. (1997). GRIP: a synaptic PDZ domain-containing protein that interacts with AMPA receptors. Nature *386*, 279-264.

Doyle, D. A., Lee, A., Lewis, J., Kim, E.. Sheng, M., and MacKinncn, R. (1996). Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. Cell *85*,1067-1076.

Du. Y., Weed, S. A.. Xiong. W.-C., Marshall, T. D., and Parsons, J. T. (1996). Identification of a novel cortactin SH3 domain-binding protein and its localization to growth cones of cultured neurons. Mol. Cell. Biol. *18*, 5638-5651.

Dunah, A., Luo. J., Wang, Y., Yasuda, R.. and Wolfe. B. (1998). Subunit composition of N-methyl-D-aspartate receptors in the central nervous system that contain the NR2D subunit. Mol. Pharmacol. *53*,429437.

Fischer, M.. Kaech. S.. Knutti. D., and Matus, A. (1996). Rapid actin-based plasticity in dendritfc spines. Neuron *28*, 647-654.

Goslin. K.. and Banker. G. (1991). Rat hippocampal neurons in low density culture. In: Culturing Nerve Cells, G. Banker and K. Goslin, eds. (Cambridge, MA: The MIT Press).

Hsueh, Y.-P., Kim, E., and Sheng. M. (1997). Disulfide-linked head-to-head multimerization in the mechanism of ion channel clustering by PSD-95. Neuron *18*, 603-814.

Huang, C., Ni, Y., Wang, T., Gao, Y., Haudenschild, C. C., and Zhan, X. (1997). Down-regulation of the Namentous actfn cross-linking activity of cortactin by Src-mediated tyrosine phosphorylation. J. Biol. Chem. *272*, 13911-13915.

Jacobs, M. D., and Harrison, SC. (1998). Structure of an IκB /N FκB complex. Cell *95*, 749-758.

Jousset, C., Carron, C., Boureux, A., Quang, C., Oury, C.. Dusanter-Fourt. I., Charon, M.. Levin, J., Bernard, 0.. and Ghysdael. J. (1997). A domain of TEL conserved in a subset of ETS proteins defines a specifc oligomerization interface essential to the mitogenic properties of the TEL-PDGFR beta oncoprotein. EM60 J. *16*, 69-62.

Kanner, S. B.. Reynolds, A. B., Vines, R. R., and Parsons, J. T. (1990). Monoclonal antibodies to individual tyrosine-phosphorylated protein substrates of oncogene-encoded tyrosine kinases. Proc. Nab Acad. Sci. USA *87*, 3328-3332.

Kato, A., Ozawa, F., Saitoh, Y., Fukazawa, Y., Sugiyama, H., and Inokuchi. K. (1998). Novel members of the Ves/Homer family of PDZ proteins that bind metabotropic glutamate receptors. J. Biol. Chem. *273*, 23969-23975.

Kennedy, M. B. (1997). The postsynaptic density at glutamatergic synapses. Trends Neurosci. *20*, 264-268.

Kim, E.. Niethammer, M., Rothschild, A., Jan, Y. N., and Sheng. M. (1995). Clustering of shaker-type K^+ channels by interaction with a family of membrane-associated guanylate kinases. Nature *378*, 85-88.

Kim, E., Cho. K.-O., Rothschild, A., and Shenq. M. (1996). Heteromultimerzation and NMDA receptor clustering activity of chapsyn-110, a member of the PSD-95 family of proteins. Neuron *17*, 103-113.

Kim, E.. Naisbitt. S., Hsueh. Y.-P., Rao, A., Rothschild, A., Craig, A. M., and Sheng. M. (1997). GKAP, a novel synaptic protein that interacts with the guanylate kinase-like domain of the PSD-95/SAP90 family of channel clustering molecules. J. Cell Biol. *136*, 669-678.

Kim, E.. Niethammer, M., Rothschild, A., Jan, Y. N.. and Sheng, M. (1995). Clustering of shaker-type K^+ channels by interaction with a family of membrane-associated guanylate kinases. Nature *378*, 85-88.

Kornau, H., Seeburg, P., and Kennedy, M. (1997). Interaction of ion channels and receptors with PDZ domain proteins. Curr. Opin. Neurobiol. *7*, 368-373.

Muller. B. M., Kistner, Lt., Veh, R. W.. Cases-Langhoff, C.. Becker, B. Gundelfinger, E. D., and Garner, CC. (1995). Molecular characterization and spatial distribution of SAP97, a novel presynaptic protein homologous to SAP90 and the Drosophia discs-large tumor suppressor protein. J. Neurosci. *15*, 2354-2366.

Muller. B. M., Kistner. U.. Kindler, S., Chung, W. J., Kuhlendahl, S., Lau, L.-F., Veh, R. W., Huganir, R. L., Gundelfinger, E. D., and Garner, CC. (1996). SAP102, a novel postsynaptic protein that interacts with the cytoplasmic tail of the NMDA receptor subunit NR2B. Neuron *17*, 255-265.

Muller. B. M., Kistner. U.. Veh. R. W., Cases-Langhoff. C., Becker, B., Gundelfinger, E. D., and Garner, CC. (1995). Molecular characterization and spatial distribution of

SAP97, a novel presynaptic protein homologous to SAP90 and the *Drosophila* discs-large tumor suppressor protein. J. Neurosci. *15*, 2354-2366.

Naisbitt, S.. Kim, E.. Weinberg, R. J., Rao, A., Yang, F.-C., Craig, A. M., and Sheng, M. (1997). Characterization of guanylate kinase-associated protein, a postsynaptic density protein at excitatory synapses that interacts directly with postysynaptic density-95/synapse-associated protein 90. J. Neurosci. *17*, 5687-5696.

Nguyen, J. T., Turck. C. W., Cohen, F. E., Zuckermann. R. N., Lim, W. A. (1998). Exploiting the basis of proline recognition by SH3 and WW domains: design of N-substituted inhibitors. Science *282*, 2088 2092.

Niethammer, M.. and Sheng, M. (1998). Identification of ion channel-associated proteins using the yeast two-hybrid system. In: Methods in Enzymology 293, P. M. Conn, ed. (New York: Academic Press), pp. 104-122.

O'Brien, R.. Lau, L., and Huganir, R. (1998). Molecular mechanisms of glutamate receptor clustering at excitatory synapses. Curr. Opin. Neurobiol. *8*, 364-369.

Peterson, A. J., Kyba, M., Bomemann, D., Morgan, K., Brock, H., and Simon, J. (1997). A domain shared by the Polycomb group proteins Scm and ph mediates heterotypic and homotypic interactions. Mol. Cell. Biol. *17*, 6683-6692.

Phend. K. D., Rustioni, A., and Weinberg, R. J. (1995). An osmium-free method of Epon embedment that preserves both ultrastructure and antigenicity for postembedding immunocytochemistry. J. Histochem. Cytochem. *43*, 283-292.

Rao, A., Kim, E., Sheng. M., and Craig, A. (1998). Heterogeneity in the molecular composition of excitatory postsynaptic sites during development of hippocampal neurons in culture. J. Neurosci. *18*, 1217-1229.

Satoh, K., Yanai, H.. Senda, T.. Kohu, K., Nakamura. T.. Okumura. N.. Matsumine. A., Kobayashi, S.. Toyoshima, K.. and Akiyama, T. (1997). DAP-1. a novel protein that interacts with the guanylate kinase like domains of hDLG and PSD-95. Genes Cells *2*, 415-424.

Sheng. M. (1996). PDZs and receptor/channel clustering: rounding up the latest suspects. Neuron *17*, 575-578.

Sheng, M. (1997). Glutamate receptors put in their place. Nature *386*, 221-223.

Songyang, Z., Fanning. AS., Fu, C., Xu, J., Marfatia. SM., Chishti, A. H., Crompton, A., Chan, AC., Anderson, J. M., and Cantley. L. C. (1997). Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. Science *275*. 73-77.

Sparks, A., Rider, J., Hoffman, N., Fowlkes, D., Quillam. L., and Kay, B. (1996). Distinct ligand preferences of Src homology 3 domains from Src, Abl. cortactin, p53BP2. PLCγ Crk. and Grb2. Proc. Natl. Acad. Sci. USA *93*, 1540-1544.

Srivastava. S., Osten, P., Vilim, F.. Khatd. L., Inman. G., States, 6.. Daly, C., DeSouza, S., Abagyan. R., Valtschanoff, J., Weinberg, R., and Ziff, E. (1998). Novel anchorage of GluR2/3 to the postsynaptic density by the AMPA receptor-binding protein ABP. Neuron *21*. 581-591.

Takeuchi, M., Hata, Y., Hirao. K., Toyoda, A., Me. 54.. and Takai, Y. (1997). SAPAPs. a family of PSD-95/SAP9O-associated proteins localized at postsynaptic density. J. Biol. Chem. *272*, 11943-1 1951.

Thanos, CD., Goodwill, K. E., and Bowie. J. U. (1999). Oligomeric structure of the human EphB2 receptor SAM domain. Science *283*, 833-836.

Tu, J. C., Xiao, 6.. Yuan. J. P., Lanahan, A. A.. Leoffert, K., Li, M.. Linden, D. J., and Worley, P. F. (1998). Homer binds a novel proline-rich motif and links group 1 metabotropic glutamate receptors with IP3 receptors. Neuron *27*, 717-726.

Tu, J. C., Xiao, 6.. Naisbitt. S., Yuan. J. P.. Petralia, R. S., Brakeman, P.. Doan, A., Aakalu. V. K., Lanahan, A. A., Sheng, M.. and Worley. P. F. (1999). Coupling of mGluR/Homer and PSD-95 complexes by the Shank family of postsynaptic density proteins. Neuron *23*, 583-592.

Weed, S. A.. Du. Y.. and Parsons, J. T. (1998). Translocation of cortactin to the cell periphery is mediated by the small GTPase Rac1. J. Cell Sci. *111*, 2433-2443.

Wenthold, R. J., Yokotani, N.. Doi, K., and Wada, K. (1992). Immunochemical characterization of the non-NMDA glutamate receptor using subunit-specific antibodies. Evidence for a hetero-oligomeric structure in rat brain. J. Biol. Chem. *267*, 501-507.

Wenthold. R. J.. Petralia. R. S.. Blahos, J. I., and Niedzielski, AS. (1996). Evidence for multiple AMPA receptor complexes in hippocampal CAl/CA2 neurons. J. Neurosci. *16*, 1982-1989.

Wu, H.. and Parsons, J. T. (1993). Cortactin, and 80/85-kilodalton pp60src substrate, is a filamentous actin-binding protein enriched in the cell cortex. J. Cell Biol. *120*, 1417-1426.

Wyszynski. M.. Kim, E., Yang, F.-C., and Sheng, M. (1998). Biochemical and immunocytochemical characterization of GRIP, a putative AMPA receptor anchoring protein, in rat brain. Neuropharmacology *37*, 1335-1334.

Xia, Z.. Dudek. H.. Miranti. C., and Greenberg, M. (1996). Calcium influx via the NMDA receptor induces immediate earfy gene transcription by a MAP kinase/ERK-dependent mechanism. J. Neurosci. *16*, 5425-5436.

Xia, J.. Zhang, X., Staudinger, J., and Huganir, R. L. (1999). Clustering of AMPA receptors by the synaptic PDZ domain-containing protein PICK1. Neuron *22*, 179187.

Xiao, 6.. Tu, J. C., Petralia. R. S., Yuan, J. P., Doan, A., Breder. CD.. Ruggiero. A., Lanahan. A. A.. Wenthold. R. J., and Worley, P. F. (1998). Homer regulates the association of group 1 metabotropic glutamate receptors with multivalent complexes of homer-related, synaptic proteins. Neuron *21*, 707-716.

Ziff, E. B. (1997). Enlightening the postsynaptic density. Neuron *19*, 1163-1174.

It is to be understood that while the invention has been described in conjunction with the detailed description thereof, the foregoing description is intended to illustrate and not limit the scope of the invention, which is defined by the scope of the appended claims. Other aspects, advantages, and modifications are within the scope of the following claims.

What is claimed is:

1. A substantially pure polypeptide characterized as:

(a) having an ankyrin domain;

(b) having an SH3 domain;

(c) having a PDZ domain;

(d) having a proline-rich domain; and

(e) having a SAM domain,

and conservative variants thereof.

2. The polypeptide of claim 1, wherein the polypeptide has an expression pattern in brain tissue.

3. The polypeptide of claim 1, wherein the polypeptide interacts with an intracellular protein selected from the group consisting of a cortactin protein, a PSD-95 protein, a Homer protein, a GKAP protein, and any combination thereof.

4. The polypeptide of claim 1, wherein the polypeptide has an amino acid sequence as set forth in SEQ ID NO:2, SEQ ID NO:4 or SEQ ID NO:6.

5. A substantially pure polypeptide having an amino acid sequence as set forth in SEQ ID NO: 2, SEQ ID NO:4 or SEQ ID NO:6, or conservative variants thereof.

6. An isolated polynucleotide encoding a polypeptide as in claim 1.

7. An isolated polynucleotide selected from the group consisting of:

(a) a polynucleotide encoding a polypeptide having an amino acid sequence as set forth in SEQ ID NO:2 or SEQ ID NO:4;

(b) a polynucleotide of (a), wherein T can be U;

(c) a polynucleotide complementary to (a) or (b);

(d) a polynucleotide having a nucleotide sequence as set forth in SEQ ID NO:1 or SEQ ID NO:3;

(e) degenerate variants of (a), (b), (c) or (d);

(f) a fragment of (a), (b), (c), (d) or (e) having at least 15 base pairs and that hybridizes to a polynucleotide encoding a polypeptide as set forth in SEQ ID NO:2 or SEQ ID NO:4; and

(g) a fragment of (a), (b), (c) (d) or (e) having at least 15 base pairs and that hybridizes to a polynucleotide encoding a polypeptide as set forth in amino acid residues 1 to 552 of SEQ ID NO:2 or residues 1 to 540 of SEQ ID NO:4.

8. An isolated polynucleotide, wherein the nucleotide is at least 15 bases in length which hybridizes under moderately to highly stringent conditions to DNA encoding a polypeptide as set forth in SEQ ID NO:2 or SEQ ID NO:4 or SEQ ID NO:6.

9. An antibody that binds to a polypeptide of claim 1 or claim 5 or binds to immunoreactive fragments thereof.

10. The antibody of claim 9, wherein the antibody is polyclonal.

11. The antibody of claim 9, wherein the antibody is monoclonal.

12. An expression vector comprising a polynucleotide of claim 7.

13. The expression vector of claim 12, wherein the vector is virus-derived.

14. The expression vector of claim 12, wherein the vector is plasmid-derived.

15. A host cell comprising a vector of claim 12.

16. A method for producing a polypeptide comprising the steps of:
    (a) culturing a host cell of claim 15 under conditions suitable for the expression of the polypeptide; and
    (b) recovering the polypeptide from the host cell culture.

17. A transgenic non-human animal having a transgene that expresses a polypeptide of claim 1 chromosomally integrated into the germ cells of the animal.

18. The transgenic animal of claim 17, wherein the animal is murine.

19. A substantially pure polypeptide, wherein the polypeptide has a PDZ domain and interacts with amino acid sequence -X-T/S-R/K-L*, wherein X is any amino acid and L* is a carboxyl-terminal leucine residue.

20. The polypeptide of claim 19, wherein the amino acid sequence is -Q-T-R-L*.

21. A polynucleotide encoding the polypeptide of claim 19.

22. A computer readable medium having stored thereon a nucleic acid sequence selected from the group consisting of SEQ ID NO:1, SEQ ID NO:3, SEQ ID NO:5, and sequences substantially identical thereto, or a polypeptide sequence selected from the group consisting of SEQ ID NO:2, SEQ ID NO:4, SEQ ID NO:6 and sequences substantially identical thereto.

23. A computer system comprising a processor and a data storage device wherein said data storage device has stored thereon a nucleic acid sequence selected from the group consisting of SEQ ID NO:1, SEQ ID NO:3, and sequences substantially identical thereto, or a polypeptide sequence selected from the group consisting SEQ ID NO:2, SEQ ID NO:4, and sequences substantially identical thereto.

24. The computer system of claim 23, further comprising a sequence comparison algorithm and a data storage device having at least one reference sequence stored thereon.

25. The computer system of claim 24, wherein the sequence comparison algorithm comprises a computer program which indicates polymorphisms.

26. The computer system of claim 23, further comprising an identifier which identifies features in said sequence.

27. A method for comparing a first sequence to a reference sequence wherein said first sequence is a nucleic acid sequence selected from the group consisting SEQ ID NO:1, SEQ ID NO:3, SEQ ID NO:5, and sequences substantially identical thereto, or a polypeptide sequence selected from the group consisting of SEQ ID NO:2, SEQ ID NO:4, SEQ ID NO:6, and sequences substantially identical thereto comprising:

(a) reading the first sequence and the reference sequence through use of a computer program which compares sequences; and

(b) determining differences between the first sequence and the reference sequence with the computer program.

28. The method of claim 27, wherein determining differences between the first sequence and the reference sequence comprises identifying polymorphisms.

29. A method for identifying a feature in a sequence wherein the sequence is selected from the group consisting of a nucleic acid sequence SEQ ID NO:1, SEQ ID NO:3, sequences substantially identical thereto, or a polypeptide sequence SEQ ID NO:2, SEQ ID NO:4, and sequences substantially identical thereto comprising:

(a) reading the sequence through the use of a computer program which identifies features in sequences; and

(b) identifying features in the sequences with the computer program.

30. A method for identifying a compound that modulates a cellular response mediated by a Shank protein comprising:

(a) incubating the compound and a cell expressing a Shank protein under conditions sufficient to permit the compound to interact with the cell;

(b) exposing the cell to conditions that activate the Shank protein; and

(c) comparing a cellular response in the cell incubated with the compound with the cellular response of a cell not incubated with the compound, thereby identifying a compound that modulates a cellular response.

31. The method of claim 30, wherein the Shank protein is Shank1a, or Shank3a, or Shank 2.

32. The method of claim 30, wherein the Shank protein is Shank1a.

33. The method of claim 30, wherein the Shank protein is Shank3a.

34. The method of claim 30, wherein the cellular response is an increase in cellular cytoskeletal stability.

35. The method of claim 30, wherein the cellular response is a decrease in cellular cytoskeletal stability.

36. The method of claim 30, wherein the cell further expresses at least one intracellular protein that interacts with a Shank protein.

37. The method of claim 36, wherein the intracellular protein is a GKAP protein, a PDS-95 protein, a cortactin protein, a Homer protein, or any combination thereof.

38. The method of claim 30, wherein the compound is selected from the species consisting of a peptide, a peptidomimetic, a polypeptide, a pharmaceutical, a chemical compound, a biological agent, an antibody and a neurotropic agent.

39. A method for identifying a compound that modulates cytoskeletal stability comprising:

(a) incubating the compound and a cell expressing a Shank protein under conditions sufficient to permit the compound to interact with the cell;

(b) exposing the cell to conditions sufficient to affect cytoskeletal stability; and

(c) comparing the cytoskeletal stability in the cell incubated with the compound with the cytoskeletal stability of a cell not incubated with the compound, thereby identifying a compound that modulates cytoskeletal stability.

40. The method of claim 39, wherein the Shank protein is Shank1a, Shank2 or Shank3a.

41. The method of claim 39, wherein the Shank protein is Shank1a.

42. The method of claim 39, wherein the Shank protein is Shank 3a.

43. The method of claim 39, wherein the modulation of cytoskeletal stability is an increase in cytoskeletal stability.

44. The method of claim 39, wherein the modulation of cytoskeletal stability is a decrease in cytoskeletal stability.

45. The method of claim 39, wherein the compound is selected from the species consisting of a peptide, a peptidomimetic, a polypeptide, a pharmaceutical, a chemical compound, a biological agent, an antibody and a neurotropic agent.

46. The method of claim 39, wherein the cell further expresses at least one intracellular protein that interacts with a Shank protein.

47. The method of claim 46, wherein the intracellular protein is a GKAP protein, a PSD-95 protein, a cortactin protein, a Homer protein, or any combination thereof.

48. A method for identifying a compound that modulates receptor localization comprising:

(a) incubating the compound and a cell expressing a Shank protein under conditions sufficient to permit the compound to interact with the cell;

(b) exposing the cell to conditions sufficient to affect receptor localization; and

(c) comparing the receptor localization in the cell incubated with the compound with the receptor localization of a cell not incubated with the compound, thereby identifying a compound that modulates receptor localization.

49. The method of claim 48, wherein the Shank protein is Shank1a, Shank2 or Shank3a.

50. The method of claim 48, wherein the Shank protein is Shank1a.

51. The method of claim 48, wherein the Shank protein is Shank 3a.

52. The method of claim 48, wherein the receptor is a cell surface receptor.

53. The method of claim 52, wherein the receptor is a glutamate receptor.

54. The method of claim 53, wherein the receptor is an ionotropic glutamate receptor.

55. The method of claim 54, wherein the receptor is a NMDA receptor.

56. The method of claim 48, wherein the modulation of receptor localization is an increase in receptor synaptic clustering.

57. The method of claim 48, wherein the modulation of receptor localization is a decrease in receptor synaptic clustering.

58. The method of claim 48, wherein the compound is selected from the species consisting of a peptide, a peptidomimetic, a polypeptide, a pharmaceutical, a chemical compound, a biological agent, an antibody and a neurotropic agent.

59. The method of claim 48, wherein the cell further expresses at least one intracellular protein that interacts with a Shank protein.

60. The method of claim 59, wherein the second intracellular protein is a GKAP protein, a PSD-95 protein, a cortactin protein, a Homer protein, or any combination thereof.

61. A method of identifying a compound that inhibits Shank protein activity comprising:

(a) designing a potential inhibitor for Shank protein activity that will form non-covalent bonds with amino acids in a Shank protein binding site based upon the crystal structure co-ordinates of Shank protein binding domain;

(b) synthesizing the inhibitor; and

(c) determining whether the inhibitor inhibits Shank protein activity.

62. The method of claim 61, wherein the Shank protein activity is stimulation of receptor synaptic clustering.

63. A method for identifying a compound that affects the formation of cell surface receptors into clusters, comprising:

(a) incubating the compound and a cell expressing a Shank protein and a Homer protein under conditions sufficient to allow the compound to interact with the cell;

(b) determining the effect of the compound on the formation of cell-surface receptors into clusters; and

(c) comparing the formation of cell-surface receptors into clusters of the cell contacted with the compound with the formation of cell-surface receptors into clusters in a cell not contacted with the compound, thereby identifying a compound that affects the formation of cell-surface receptors into clusters.

64. The method of claim 63, wherein the cell-surface receptor is a NMDA receptor.

65. The method of claim 63, wherein the cell-surface receptor is a metabotropic glutamate receptor.

66. The method of claim 65, wherein the metabotropic glutamate receptor is a group I metabotropic glutamate receptor.

67. The method of claim 66, wherein the metabotropic glutamate receptor is metabotropic glutamate receptor 1α.

68. The method of claim 66, wherein the metabotropic glutamate receptor is metabotropic glutamate receptor 5.

69. The method of claim 63, wherein the Shank protein is Shank1a, Shank2 or Shank3a.

70. The method of claim 63, wherein the Homer protein is Homer 1a, Homer 1b, Homer 1c, Homer 2a, Homer 2b and Homer 3.

71. The method of claim 63, wherein the compound is selected from the species consisting of a peptide, a peptidomimetic, a polypeptide, a pharmaceutical, a chemical compound, a biological agent, an antibody and a neurotropic agent.

72. The method of claim 63, wherein the effect is inhibition of the recruitment of cell-surface receptors into clusters.

73. The method of claim 63, wherein the effect is stimulation of the formation of cell-surface receptors into clusters.

74. The method of claim 63, wherein said cell is selected from the group consisting of a neuronal cell, a glial cell, a cardiac cell, a bronchial cell, a uterine cell, a testicular cell, a liver cell, a renal cell, an intestinal cell, a thymus cell, a spleen cell, a placental cell, a skeletal muscle cell and a smooth muscle cell.

75. A method of treating a disorder associated with glutamate receptors comprising administering to a subject in need thereof a therapeutically effective amount of a compound that modulates a Shank protein activity.

76. The method of claim 75, wherein the Shank protein is Shank1a, Shank2 or Shank3a.

77. The method of claim 75, wherein the disorder is selected from epilepsy, glutamate toxicity, disorders of memory, disorders of learning, stroke, schizophrenia, Alzheimer's disease, tissue degeneration and disorders of brain development.

78. A method of treating a disorder associated with a Shank protein activity comprising administering to a subject in need thereof a therapeutically effective amount of a compound that modulates a Shank protein activity.

79. The method of claim 78, wherein said disorder is a cardiac disorder, a disorder of musculature, a vasculature disorder, a neurological disorder, a psychiatric disorder, a renal disorder, a uterine disorder or a disorder of bronchial tissue.

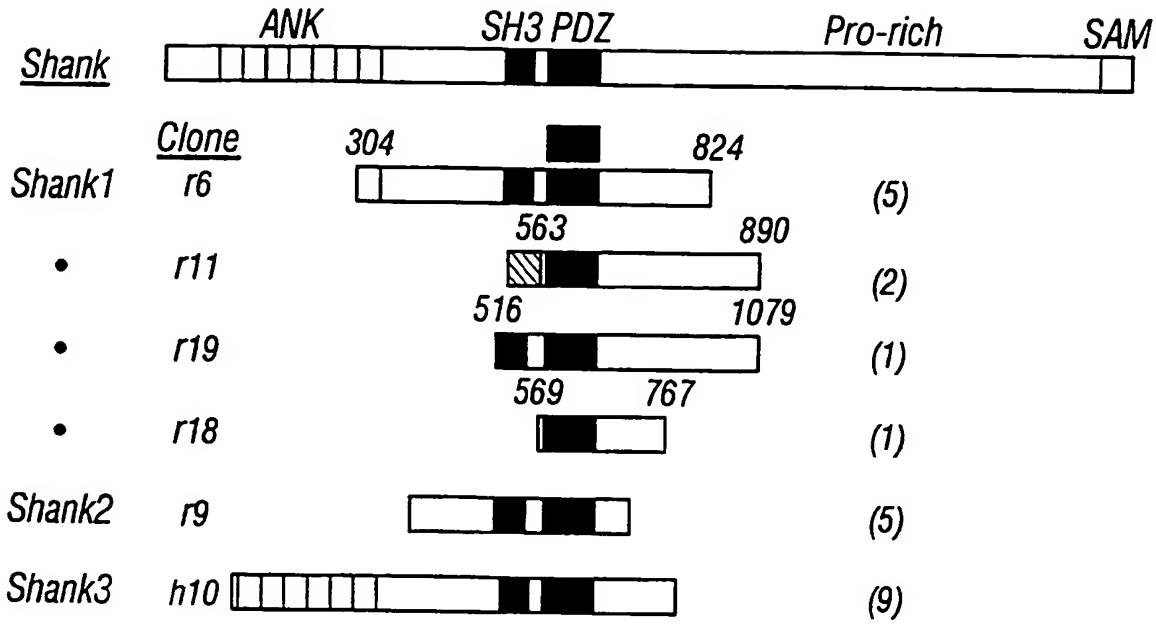

FIG. 1A

LexA \ GAD	Shank1(SH3-PDZ)	Shank1(PDZ)	PSD95(PDZ1-2)	PSD95(GK)
GKAP1a(591-666)	+++/+++	++/++	-/-	-/-
GKAP1a(660-666)	+++/+++	+++/+++	-/-	-/-
GKAP1a(1-100)	-/-	-/-	-/-	+++/+++
GKAP1b(602-627)	-/-	-/-	-/-	-/-
Kv1.4(568-655)	-/-	-/-	+++/+++	-/-

FIG. 1B

GKAP1a C-terminus				β-GAL/HSS
(wt) Q	T	R	L	+++/+++
Q	T	R	V	+/++
Q	T	R	A	-/-
Q	T	K	L	++/+++
Q	T	Q	L	-/+
Q	S	R	L	+/+++
Q	A	R	L	-/-
E	T	R	L	+++/+++
A	T	R	L	+++/+++

FIG. 1C

```
Shank1a    (1) --------MV FRIGIPDLHQ TKCLRFNPDA TIWTAKQQVL CALSESLQDV LNYGLFQPAT SGRDANFLEE
Shank3a    (1) MDGPGASAVV VRVGIPDLQQ TKCLRLDPTA PVWAAKQRVL CALNHSLQDA LNYGLFQPPS RGRAGKFLDE
                                                                  r1
Shank1a   (63) ERLLREYPQS FEXGVPYLEF RYKTRVYKQT NLDEKQLAKL HTKTGLKKFL EYVQLGTSKD VARLLDKGLD
Shank3a   (71) ERLLQDYPPN LDTPLPYLEF RYKRRVYAQN LIDDKQFAKL HTKANLKKFM DYVQLHSTDK VARLLDKGLD
                     r2                                           r3
Shank1a  (133) PNYHDSDSGE TPLTLAAQTE GSVEVIRTLC LGGAHIDFRA RDGMTALHKA ACARHCLALT ALLDLGGSPN
Shank3a  (141) PNFHDPDSGE CPLSLAAQLD NATDLLKVLR NGGAHLDFRT RDGLTAVHCA TQRNAGALT TLLDLGASPD
                  r4                                           r5
Shank1a  (203) YKDRRGLTPL FHTAMVGGDP RCCELLLYNR AQLGIADENG WQEIHQACQR GHSQHLEHLL FYGAEPGAQN
Shank3a  (211) YKDSRGLTPL YHSALGGGDA LCCELLLHDH AQLGTTDENG WQEIHQACRF GHVQHLEHLL FYGANMGAQN
                    r6                                         r7
Shank1a  (273) ASGNTALHIC ALYNKETCAR ILLYRGANKD VKNNNGQTPF QVAVIAGNFE LGELIRNHRE QDVVPFQESP
Shank3a  (281) ASGNTALHIC ALYNQESCAR VLLFRGANKD VRNYNSQTAF QVAIIAGNFE LAEVIKTHKD SDVVPFRETP

Shank1a  (343) KYAARRRGPP GAGLTVPPAL LRANSDTSMA LPDWMVFSAP GASSSGTPGP TSGPQGQSQP SAPSTKLSSG
Shank3a  (351) SYAKRRRLAG PSGLASPRPL QRSASDINL. .........K GDQPAASPGP T.......LR SLPHQLLLQR

Shank1a  (413) TLRSASSPRG ARARSPSRGR HPEDAKRQPR GRPSSSGTPR DGPAGGTGGS GGPGGSLGSR GRRRKLYSAV
Shank3a  (404) LQEEKDRDRD GEQENDISG. ..PSAGRGGH SKISPSGPGG SGPAPGPGPA SPAPPAPPPR GPKRKLYSAV
```

ANK

FIG. 2A

```
Shank1a   (483) PGRSFMAVKS YQAQGEGEIS LSKGEKIKVL SIGEGGFWEG QVKGRVGWFF SDCLEEVANR SQEGRQESRS
Shank3a   (471) PGRKFIAVKA HSPQGEGEIP LHRGEAVKVL SIGEGGFWFG TVKGRTGWFP ADCVEEVQMR QYDTRHETRE

Shank1a   (553) DKAKRLFRHY TVGSYDSFDA PSLIDGIDSG SDYIIKEKTV LLQKKDSEGF GFVLRGAKAQ TPIEEFTPTP
Shank3a   (541) DRTKRLFRHY TVGSYDS... ......LTSH SDYVIDDKVA ILQKRDHEGF GFVLRGAKAE TPIEEFTPTP

Shank1a   (623) AFPALQYLES VDEGGVAWRA GLRMGDFLIE VNGQNVVKVG HRQVVNMIRQ GGNTLMVKVV MVTRHPDMDE
Shank3a   (602) AFPALQYLES VDVEGVAWKA GLRTGDFLIE VNGVNVVKVG HKQVVGLIRQ GGNRLVMKVV SVTRKPEEDS

Shank1a   (693) AVHKKASQQA KRLPPPAISL RSKSMTSELE EMEYEQQPAA VPSMEKKRTV YQMALNKLDE ILAAAQQTIS
Shank3a   (672) A.RRRAPPPP KRAPSTTLTL RSKSMTAELE EL..... .....ASI RRRKGEKLDE ILAVAAEPTL

Shank1a   (763) ASESPGPGGL ASLGKHRPKG FFATESSFDP HHRSQPSYDR PSFLPPGPGL ML.RQKSIGA AEDDRPYLAP
Shank3a   (726) RPDIADADSR AATVKQRPTS RRITPAEISS LFERQ.GLPG PEKLPGSLRK GIPRTKSVGE DEKLASLLEG

Shank1a   (832) P.AMKFSRSL SVPGSEDIPP PPTTSPPEPP YSTPPAPSSS GRLTPSPRGG PFNPSSGGPL PASSPSSFDG
Shank3a   (795) RFPRSTSMQD TVREGRGIPP PPQTAPPPPP ...APYYFDS G...PPPTFS PPPPPPGRAY DTVRSSFKPG

Shank1a   (901) PSPPDTRGGG REKSLYHSAA LPPAHHHPPH HHHHHAPPPQ PHHHHAHPPH PPEMETGGSP DDPPPRLALG
Shank3a   (859) ...LEARLGA GAAGLYDSGT ..... ..... PLGPLPYPER QKRARSMIIL QDSAPEV...
```

SH3

PDZ

FIG. 2A-2

```
Shank1a  (971) PQPSLRGWRG GGPSPTSGAP SPSHHSSSGG SSGPTQAPAL RYFQLPPRAA SAAMYVPARS GRGRKGPLVK
Shank3a  (903) .........G DVPRPAPAAT PPERPKRRPR PSGPDSP... .YANL..GAF SASLFAPSKP QR.RKSPLVK

Shank1a (1041) QTKVEGEPQK GSIPSASSPT SPALPRSEPP PAGPSEKNSI PIPTIIIKAP STSNSGRSSQ GSSTEAEPPT
Shank3a  (957) QLQVEDAQER A......... ..ALAVGSPG PVGGS..... .....FAREP SPTHRGPRPG GLDY.....S

Shank1a (1111) QPDGAGGGGS SPSPAPATSP VPPSPSPVPT PASPSGPATL DFTSQFGAAL VGAARREGGW QNEARRRSTL
Shank3a (1001) SGEGLGLTFG GPSPGPVKER .......... .......... .......... .......... RLEERRRSTV

Shank1a (1181) FLSTDAGDED GGDSGLGPGG PPGPRLRHSK SIDEGMFSAE PYLRLESGGS SGGYGAYAAG SRAYGGSGSS
Shank3a (1031) FLSVGAIEGN PPSADL.... ...PSLQPSR SIDERLLGT. ...GATTGRD LLLPSPVSAL KPLVGGPNLG

Shank1a (1251) SAFTSFLPPR PLVHPLTGKA LDPASPLGLA LAARERALKE SSEGGGTPQP PPRPPSPRYD APPPTLHHHS
Shank3a (1090) PSSSTF.... ..IHPLTGKP LDPSSPLALA LAARERAL.. ......ASQT PSRSPTP... ..........

Shank1a (1321) PHSPHSPHAR HEPVLRLWGD PARRELGYRA GLGSQEKALT ASPPAARRSL LHRLPPTAPG VGPLLLQLGP
Shank3a (1133) ...VHSPDAD RPGPLFVDVQ TRDSERGPLA SPAFSPRSPA WIPVPARR.. .EAEKPTR.. .........E

Shank1a (1391) EPPTPHPGVS KAWRTAAPEE PERLPLHVRF LENCQARPPP AGTRGSSTED GPGVPPPSPR RVLPTSPTSP
Shank3a (1186) ERKSPEDKKS MILSVLDTSL QRPAGLIVVH ATSNGQEPNR LGAEEERPGT PELAPTPMQA AAVAEPMPSP
```

FIG. 2A-3

```
Shank1a  (1461) RGNEENGLPL LVLPPPAPSV DVDDGEFLFA EPLPPPLEFS NSFEKPESPL TPGPPHPLPD PPSPATPLPA ← Homer
Shank3a  (1256) RAQPPGNIP. .ADPGPSQGN SEEEPKLVFA VNLPPAQLSS NDEETREELA RIGLVPPPEE FANGI..LLA ← binding

Shank1a  (1531) APPPAVAAAP PTLDSTASSL TSYDSEVATL TQGAPAAPGD PPAPGPPAPA APAPPAPQPG PDPPPGTDSG
Shank3a  (1322) TPPPGPGPLP TTVPSPAS.. ..........  ........GK PSSELPPAPE SAA.......  .......DSG

Shank1a  (1601) IEEVDSRSSS DHPLETISSA STLSSLSAEG GGNTGGVAGG GAGVANGTEL LDTYVAYLDG QAFGGSGTPG
Shank3a  (1358) VEEADTRSSS DPHLETTSTI STVSSMSTLS ..........  ....SESGEL TDTHTSFADG HTFLLEKPPV ← Cortactin
                                                                                              binding
Shank1a  (1671) PPYPPQLMTP SKLRGRALGT SGNLRPGPSG GLRDPVTPTS PTVSVTGAGT DGLLALSACP GPSTAGVAGG
Shank3a  (1414) PP.KPKLKSP .......LG. ..........  ..KGPVTFRG PLLK...QSS DSELMAQQHH ATST....GL

Shank1a  (1741) PVAVEPEVPP VPLPAASSLP RKLLPWEEGP GPPPPPLPGP LSQPQASALA TVKASIISEL SSKLQQFGGS
Shank3a  (1456) TSAAGPARPR YLFQRRSKL. .....W..GD PVESRGLPGP EDD.......  ..KPTVISEL SSRLQQLNKD

Shank1a  (1811) STAGGALPWA RGGSGGSTDS HHGGASYIPE RTSSLQRQRL SEDSQTSLLS KPSSSIFQNW PKPPLPPLPT
Shank3a  (1509) TRSLGEEP.. VGGLGSLLD. ........PA KKSPIAAARC A........V VPSA....GW LFSSLGELST

Shank1a  (1881) GSGVSSSTAA APGATSPSAS SASASTRHLQ GVEFEMRPPL LRRAPSPSLL PASDHKVSPA PRPSSLPILP
Shank3a  (1556) .....ISAQR SPGGPGGGAS .........Y SVRPSGRYPV ARRAPSP.VK PASLERVEGL GAGVGGAGRP
```

FIG. 2A-4

```
Shank1a  (1951) SGPIYPGLFD IRSSPTGGAG GSTDPFAPVF VPPHPGISGG LGGALSGASR SLSPTRLLSL PPDKPFGAKP
Shank3a  (1611) FGLTPPTILK SSSLSI.... .PHEPKEVRF VV..RSASAR SRSPSPSPLP SPSPGSGPSA GPRRPFQQKP

                                                                                    SAM
Shank1a  (2021) LGFWTKFDVA DWLEWLGLSE HRAQFLDHEI DGSHLPALTK EDYVDLGVTR VGHRMNIDRA LKFFLER*
Shank3a  (1674) LQLWSKFDVG DWLESIHLGE HRDRFEDHEI EGAHLPALTK EDFVELGVTR VGHRMNIERA LRQLDGS*
```

FIG. 2A-5

full length sequence of Shank2 (also known as ProSAP1)

```
   1 MKSLLNAFTK KEVPFREAPA YSNRRRRPPN TLAAPRVLLR SNSDNNLNAG APEWAVCSAA
  61 TSHRSLSPQL LQQTPSKPDG ATKSLGSYAP GPRSRSPSLN RLGGAGEDGK RPQPPHWHVG
 121 SPFTPGANKD SLSTFEYPGP RRKLYSAVPG RLFVAIKPYQ PQVDGEIPLH RGDRVKVLSI
 181 GEGGFWEGSA RGHIGWFPAE CVEEVQCKPR DSQAETRADR SKKLFRHYTV GSYDSFDAAS
 241 DCIIEDKTVV LQKKDNEGFG FVLRGAKADT PIEEFTPTPA FPALQYLESV DEGGVAWQAG
 301 LRTGDFLIEV NNENVVKVGH RQVVNMIRQG GNHLVLKVVT VTRNLDPDDT ARKKAPPPPK
 361 RAPTTALTLR SKSMTAELEE LVDKASVRKK KDKPEEIVPA SKPSRTAENV AIESRVATIK
 421 QRPTSRCFPA ASDVNSVYER QGIAVMTPTV PGSPKGPFLG LPRGTMRRQK SIDSRIFLSG
 481 ITEEERQFLA PPMLKFTRSL SMPDTSEDIP PPPQSVPPSP PPPSPTTYNC PRSPTPRVYG
 541 TIKPAFNQNP VAKVPPATRS DTVATMMREK GMFYRRELDR FSLDSEDVYS RSPAPQAAFR
 601 TKRGQMPENP YSEVGKIASK AVYVPAKPAR RKGMLVKQSN VEDSPEKTCS IPIPTIIVKE
 661 PSTSSSGKSS QGSSMEIDPQ ATEPGQLRPD DSLTVSSPFA AAIAGAVRDR EKRLEARRNS
 721 PAFLSTDLGD EDVGLGPPAP RMQPSKFPEE GGFGDEDETE QPLLPTPGAA PRELENHFLG
 781 GGEAGAQGEA GGPLSSTSKA KGPESGPAAA LKSSSPASPE NYVHPLTGRL LDPSSPLALA
 841 LSARDRAMQE SQQGHKGEAP KADLNKPLYI DTKMRPSVES GFPPVTRQNT RGPLRRQETE
 901 NKYETDLSKD RRADDKKNML INIVDTAQQK SAGLLMVHTV DIPVAGPPLE EEEDREDGDT
 961 KPDHSPSTVP EGVPKTEGAL QISAAPEPAA APGRTIVAAG SVEEAVILPF RIPPPPLASV
1021 DLDEDFLFTE PLPPPLEFAN SFDIPDDRAA SVPALADLVK QKKSDTPQPP SLNSSQPANS
```

FIG. 2B-1

```
1081 TDSKKPAGIS NCLPSSFLPP PESFDAVTDS GIEEVDSRSS SDHHLETTST ISTVSSISTL
1141 SSEGGESMDT CTVYADGQAF VVDKPPVPPK PKMKPIVHKS NALYQDTLPE EDTDGFVIPP
1201 PAPPPPPGSA QAGVAKVIQP RTSKLWGDVT EVKSPILSGP KANVISELNS ILQQMNRGKS
1261 VKPGEGLELP VGAKSANLAP RSPEVMSTVS GTRSTTVTFT VRPGTSQPIT LQSRPPDYES
1321 RTSGPRRAPS PVVSPTELSK EILPTPPSAA AASPSPTLSD VFSLPSQSPA GDLFGLNPAG
1381 RSRSPSPSIL QQPISNKPFT TKPVHLWTKP DVADWLESLN LGEHKETFMD NEIDGSHLPN
1441 LQKEDLIDLG VTRVGHRMNI ERALKQLLDR
```

PARTIAL CLONE OF SHANK 2:

```
   1 RSNSDNNLNA GAPEWAVCSA ATSHRSLSPQ LLQQTPSKPD GATKSLGSYA PGPRSRSPSL
  61 NRLGGAGEDG KRPQPPHWHV GSPFTPGANK DSLSTFEYPG PRRKLYSAVP GRFFVAIKPY
 121 QPQVDGEIPL HRGDRVKVLS IGEGGFWEGS ARGHIGWFPA ECVEEVQCKP RDSQAETRAD
 181 RSKKLFRHYT VGSYDSFDAA SDCIIEDKTV VLQKKDNEGF GFVLRGAKAD TPIEEFTPTP
 241 AFPALQYLES VDEGGVAWQA GLRTGDFLIE VNNENVVKVG HRQVVNMIRQ GGNHLVLKVV
 301 TVTRNLDPDD TARKKAPPPP KRAPTTALTL RSKSMTAELE ELGLSLVDKA SVR
```

FIG. 2B-2

SHANK 1 nucleotide sequence

```
   1 ggccccacga tgacccacag ccccgcgaca agcgaggacg aggaacgcca cagtgccagc
  61 gagtgtcccg agggggggctc agaatcggac agctccccag acggaccagg aagaggcccc
 121 cagggtactc ggggccgggg cagcggagca cctggtaacc tggcctctac ccgaggcctc
 181 cagggacgct ccatgtctgt ccctgacgac gcccacttca gcatgatggt tttcaggatt
 241 ggcatccctg acttgcacca gacgaaatgc ctacgcttca acccggatgc taccatctgg
 301 acagccaaac agcaagtact ctgtgccctg agcgagagtc tacaggatgt gctaaactat
 361 ggactgttcc agccagccac ctcaggccgg gatgccaact tcctggagga agagaggctt
 421 ctaagagagt accctcagtc ctttgagaaa ggggtaccct acctggagtt ccgctacaag
 481 acccgagtct acaaacagac caacctggat gagaaacagc tggccaaact gcacacgaag
 541 acggggttga agaagttcct tgaatatgtg cagctcggga cgtctgacaa ggtggcgagg
 601 ctgctggaca aggggctgga ccccaattac cacgactcag attcaggaga gacacctttg
 661 accctggcgg cccagactga aggctcagta gaggtgattc ggacgctgtg cctgggcggg
 721 gcgcacatag acttcagggc ccgggatggc atgacagcgc tgcacaaagc tgcatgtgct
 781 aggcactgcc tggcactcac ggcacttctg gaccttgggg gctcccccaa ctacaaggat
 841 cgacgagggc tgacccccct gttccacaca gctatggttg ggggagaccc ccgctgttgt
 901 gagctacttc tgtacaacag ggcccagctg ggcattgcag atgagaacgg ctggcaggaa
 961 atccaccagg cctgccaacg agggcactct cagcacctgg aacacctgct cttctatggg
1021 gctgagcctg gagcccagaa tgcctcagga aacacggccc tgcacatctg tgctctctac
```

FIG. 2C-1

```
1081 aacaaggaga cctgtgccag gatcctcctt tatcgcggag caaacaagga tgtgaaaaat
1141 aacaacggac agacaccctt ccaggtggct gtgattgctg ggaatttttga gctgggggaa
1201 ctgattcgta accatcggga acaggatgtg gtaccccttcc aggagtcccc caagtatgca
1261 gcccggcgtc ggggaccccc aggcgcaggg ctgactgtac ccccagcgct gctgcgggct
1321 aacagcgaca cgagcatggc gctgcctgat tggatggtgt ttttgggccc agtgggcctca
1381 tcctctgggga cccctggccc tacctcagga ccccagggcc agtcacagcc ctcggccccc
1441 agcaccaagc tcagtagtgg gaccctcaga agtgccagca gccccgagg tgcccgggca
1501 cgctccccat cccgggggag acatcctgaa gatgccaaga gacaacctcg aggccgaccc
1561 agctccagcg ggacaccccg agatgggccg gctgggggaa ctgggggctc agggggccct
1621 gggggctccc tgggtagcag aggcagacgt aggaagctct attcagcggt acctggacgc
1681 tccttcatgg cggtgaagtc ataccagggcg caagggcgagg gggagatctc tctgagtaaa
1741 ggcgaaaaga tcaaagtgct cagcatcggg gaaggaggct tctgggaagg ccaggtcaaa
1801 ggtcgagttg gctggttccc ttctgactgc ctagaagagg tggcaaaccg ttcccaggag
1861 ggaagacaag aaagccgaag tgacaaggca aagagactct tcaggcatta cacagtgggc
1921 tcttacgaca gcttcgatgc cccaagtctg atcgatgggga ttgactcagg gagtgattac
1981 atcatcaagg agaagacagt cttactgcag aagaaggaca gtgagggggtt cgggttcgtg
2041 ctccgggggg ccaaggcgca gacccccatc gaggaattca cccccacccc ggcccttcccg
2101 gcgctgcagt acctcgagtc tgtggacgag ggtggcgtgg catggcgggc tggactacgg
2161 atgggagact tcctcatcga ggtgaatgga cagaatgtag tgaaggttgg ccaccgccag
2221 gtggtcaaca tgatccgaca gggaggcaac acactgatgg tcaaggtggt gatggtgacc
```

FIG. 2C-2

```
2281 agacacccag acatggacga ggctgtccac aagaaagcat cccagcaggc taagcggctc
2341 ccgccccccag ccatctccct gcgttccaag tctatgacct cagagctgga ggaaatggaa
2401 tacgagcaac agcccgcggc agtgcccagc atggagaaaa agaggactgt gtatcagatg
2461 gcactcaaca aactggatga gatcttggcc gctgctcagc agaccatcag cgcaagtgag
2521 agtcctgggc ctggtggcct ggcatccctg ggcaagcacc gacccaaagg attctttgcc
2581 actgagtcga gctttgatcc tcaccaccgc tcccagccaa gttatgaccg gccttctttc
2641 ctgcctcctg gacccggcct tatgctccgg cagaaatcta tcgggggctgc ggaagacgac
2701 agaccctacc tagcacccca agccatgaaa ttcagtcgta gcctgtcagt ccccggttca
2761 gaggacatac ccccaccacc cactacatca ccaccagagc cgccctatag cacacctccg
2821 gccccctcct cctccggccg cctcactccc tcgcctcggg gagggccctt taaccccagt
2881 tctgggggtc cccttccccgc ctcctccccct tcatctttcg atgggccatc tcctcctgac
2941 acccgtggtg ggggccgaga aaagagccta tatcacagcg cagccctgcc ccctgcccac
3001 caccatccac ctcatcacca ccatcatcac gcgccacctc cccagcccca ccatcaccat
3061 gcacatcctc ctcacccacc ggagatggag acaggaggtt cccccgacga tcccccacct
3121 cgattggccc tgggacccca gcccagcctt cgaggctggc gaggtggagg acccagcccg
3181 acctcggggg ccccatcacc ttcccaccac agtagcagcg gaggcagcag cggccctact
3241 caggctcccg ccctgcgnta cttccaactg cccccgaggg cggccagtgc ggctatgtat
3301 gtaccagcac gttccggccg cggtcgcaag ggcccactgg tcaagcagac caaagtggaa
3361 ggcgagcccc aaaagggcag catcccttcc gcgtccctctc caacatcccc ggcgttgcct
3421 cggtctgaac cacctccggc cgggccgtcg gagaaaaatt ccatcccctat cccgaccatc
```

FIG. 2C-3

```
3481 atcatcaaag ccccatctac cagtaacagc ggccgcagca gtcagggaag cagcacagag
3541 gctgagcccc ccacccagcc cgatggcgcg ggcggcggcg gatcctcgcc tagccctgca
3601 cctgccacgt cccctgtgcc gccatccccg tcgcccgtgc ccactcccgc gtcacccagc
3661 ggcccggcca ccctggattt cactagccag ttggagcag ccctagtagg cgcggcgcgg
3721 agggagggag gctggcagaa cgaagcccgg cggcggtcca cgctgttcct gtccaccgac
3781 gcaggagatg aggatggagg cgacagcggg ctgggcccgg ggggcccccc gggtcccccgg
3841 ctgcgccatt ccaaatccat tgacgaaggc atgttttccg cagagccgta cctccggctg
3901 gagtcagggg gcagcagtgg gggctacggg gcctacgcag cgggcagccg agcctacggg
3961 ggcagcgggga gcagcagcgc cttcaccagc ttcctccccc cgcggccccct ggtgcatccg
4021 ctgaccggca aagccttgga tcccgcctct ccgctgcggt tggccttggc cgcccgggag
4081 cgggccttga aggagtcctc ggagggtggg ggcacccctc agccgccccc gagacctcca
4141 tcaccacgct acgacgcccc accgcctacc cttcaccacc actcgcccca ttcaccccat
4201 tcgccacatg cgcgtcacga gccagtgttg cgcctctggg gggaccccgc gcgccgggag
4261 ctgggtacc gggcaggcct agggagtcag gagaaggccc ttaccgccag tccacctgca
4321 gctcgccgct ccctactgca ccgtctgccc ccgacggctc ccggggtcgg gccgctcctg
4381 ctgcagctgg ggccggagcc cccaactccc catcctgggg tgagcaaggc ctggcgcaca
4441 gcagccccgg aggagccgga gcggctgccg ctgcacgtgc gattccttga gaattgccaa
4501 gcccggcctc ccccagctgg cacgagaggg tcctctactg aagatgggcc tgggggtcccg
4561 cctcccagcc ctcgcagagt cttgccaacc tctcccacct ccccacgagg caacgaggag
4621 aacggccttc cgctactggt cctgccaccg ccagcccctt cggtggatgt ggacgacggc
```

FIG. 2C-4

```
4681 gagttcttgt ttgcggagcc gctgcctcca cccctggagt tttccaatag tttcgaaaag
4741 ccagagtcgc ctctcacacc cgggcctcct cacccactac ctgaccctcc ctccccggct
4801 acccctctac ccgcagcccc acctcccgct gtagccgcag cccctcccac cttggactcc
4861 acagcatcca gcctgacctc ttacgacagc gaagtggcca cactgaccca ggggggctccc
4921 gcggctccctg gggatcctcc agctccaggc ccgccggccc cggcagcccc agctcccccca
4981 gcaccgcagc caggccctga tcctccacct ggcacagatt ctggaattga ggaggtggac
5041 agtcggagta gcagtgatca cccgctggag accatcagca gcgcatccac actcagcagc
5101 ttgtccgccg aaggaggagg caatacaggg ggcgtggctg ggggcggggc cggggtggcc
5161 aatgggacgg agctcctgga cacctacgtt gcctacctgg atggccaggc cttcgggggc
5221 agtggaactc ctggcccacc atatcctccc cagctcatga ctccttctaa gctccgggggc
5281 agggcgcttg ggactagtgg aaatttaaga cctggtcccca gtgggggact ccgagatcct
5341 gtcacccccca ccagccccac tgtctctgtg accggagctg gaacagatgg gctactggcc
5401 ttgagtgctt gccctggacc ctcaacggca ggtgtggcag ggggtcccgt ggcagttgag
5461 ccagaagtcc cgcctgtacc tttgcccgct gcctcatccc tgcctcggaa gctattaccg
5521 tgggaggagg gtccgggacc accaccacca ccccctgccag ggccccctgtc ccagcctcag
5581 gcctctgcct tggccacggt caaagccagc atcatcagcg aactcagctc caagctccag
5641 cagtttggtg ggtcctctac cgcaggagga gctctgcctt gggctcgggg aggcagcggg
5701 ggcagcacag acagccacca cggaggagcc agttacatcc cagagaggac gtcctccctg
5761 cagcggcaga ggctctctga agactcccag acctctctcc tctccaaacc cagcagcagc
5821 atctttcaga actggcctaa acccctctg ccaccactcc ccactgggtc tggggtctcc
```

FIG. 2C-5

```
5881 tcttcaacag ctgcagcccc gggagctaca tccccttcag cctcatctgc ctcagcttcc
5941 accagacacc tgcagggtgt ggaatttgag atgcggcctc cgctgctccg ccgggcaccc
6001 agcccctcgc tgctgcctgc ctcggatcac aaggtcagcc cagcgcccag gccttcctct
6061 ttgcccatcc tgccttcagg acccatctac ccgggcctct tcgacatccg tagctctcca
6121 accggagggg cgggaggctc gactgacccc ttcgctcccg tctttgtgcc accacacccc
6181 gggatatctg ggggtcttgg aggagccttg tctgggggcct ctcgctccct atcacccacc
6241 cgcctgctct cgctgccccc ggacaagccg ttcgggcgcca agcctctagg gttctggacc
6301 aagtttgacg tcgccgattg gctagagtgg ctgggcttgt cggagcaccg agcccagttc
6361 ctggaccacg agattgatgg ctcccatctg cctgccttga ccaaggagga ctatgtggac
6421 ctgggtgtaa ccagggtggg ccaccgtatg aacattgacc gggctctcaa attttcctg
6481 gagaggtgat ggctggcctg gacggaccag cccgtccata ggacccttga gcccgctgac
6541 ctctgacctc tgacccctct gaccatcatt ccccctctcc cctcaggggg ccagggactc
6601 tgctaaaact gtgtcctgcc ccgtctcccc aagcctgact gaggtcccca ggtggaccgg
6661 atgcaggcca gacaatggca cctcgcaaga ggtgggcagc tgacacaaga ctgcaagtgg
6721 agtagatggg ggattagaaa aaaatgacaa aggggtactg aagaaagagg gaaaaaggct
6781 gactaggggt gggggaacag acagagccat agtgggtcag tccagagcct gacctgatgg
6841 ggtcctgggg aggtgtgccc tccgattcca tcatttgcgc ctctcaaata acccttttca
6901 cctctctagg cccctcagcc ttcccactct ccttggctca ctcagctgcc cagccccctg
6961 cctcttgcac gcctgttccc ctccctggct gcagtttttg cacaagcttg ccccctggct
7021 gtcttgcaca cctccggcct gtaaggcccct ctcagcttcc caacacgctg cattcggttt
7081 aaaaaaa
```

FIG. 2C-6

SHANK 2 nucleotide sequence

```
   1 tcccggtgtc tcctggccgc cagctcgcgc gcaatcgcct gctccatgtt aaccctgcgt
  61 ggtgctgcct ccccgatgcc ctgcttgctc tcccagcaga gtggcactgc tcacatgagg
 121 atcaccgcca gaggtagccg gttctgagag cagcctcagt ctgcgcacag ttgcttccca
 181 gctgacagac agaatatggg gtgggctttg tcaccagcac actgggtctg gcatgcccag
 241 cgatgctcgg ggctgcagag cgcccagacg cctgccactg ctcgccgggt cccagccgcg
 301 ccagctccga ggggccgggc tgcgcgccta gccttcctgc ccgccgagag ggttatataa
 361 tgagcggacc gctgccaaag tcacagggaa aatgaagtct ttgttaaatg ccttcaccaa
 421 gaaggaagtg cccttccgag aggccccagc ttattccaac cgcaggcggc ggcccccaa
 481 cacattggct gcccccagag ttcttctgcg ttccaatagc gacaacaacc tcaatgccgg
 541 tgcgcctgaa tgggctgtct gctctgcagc cacctcccac cgaagcctct caccccagct
 601 gctgcagcag acacccagca aacctgatgg agccacaaag agccttggaa gctacgcccc
 661 tgggccccgc agccgctctc cctcactcaa caggctgggt ggcgctggtg aggatggcaa
 721 gaggccacag ccgccacatt ggcatgtggg gtcgcccttc actcctggtg ccaacaagga
 781 ctccctctcg acctttgagt acccagggcc ccggaggaag ctgtatagtg cggtgcctgg
 841 gagactcttc gtcgccatca agccatacca accccaagtc gacggcgaga tccccccttca
 901 ccgaggcgac agggtcaaag ttctgagcat cggcgagggc ggcttctggg aaggcagcgc
 961 ccgtggccac atcgggtggt tcccagctga gtgtgtggag gaggtacaat gtaagccccg
1021 ggacagccaa gcagaaaccc gtgcggaccg cagcaagaag ctctttcggc attacacagt
```

FIG. 2D-1

```
1081 gggctcctat gacagctttg atgctgccag tgactgcatt attgaggaca agacggtggt
1141 cctgcagaag aaagacaacg agggctttgg atttgtgctc cgagggggcaa aagctgatac
1201 ccccattgag gaattcacac ccacgccggc atttccagcc ctgcagtacc tggagtccgt
1261 ggatgaagga ggggtggcat ggcaagccgg actaaggacc ggggacttct tgattgaggt
1321 taacaatgaa aatgtcgtca aggtgggcca caggcaggtg gtgaacatga tccgccaggg
1381 agggaatcac ctcgtccttа aggtcgtcac ggtgaccagg aatctagacc ctgatgatac
1441 agccagaaag aaagctcccc cacctccaaa gcgggctccg accacggccc tcaccctgcg
1501 ttccaagtcc atgacagcgg aattggagga actcgtggat aaagcctcag tccggaagaa
1561 gaaggataaa ccggaagaga tagtcccagc ctccaagccc tccaggactg cagagaacgt
1621 ggccatagaa tccagggtgg caaccatcaa gcagcggccc acaagccggt gctttccagc
1681 tgcctctgat gtgaactccg tgtacgagcg ccaagggatt gctgtaatga cgcccacggt
1741 ccctgggagc ccgaaaggcc catttctggg cctccctcga ggtacgatgc gaaggcagaa
1801 atcgatagac agcagaatct ttctatcagg gataacagag gaagagcggc agtttctggc
1861 tcccccaatg ctgaagttca cccgaagctt gtccatgcca gacacttctg aggacatccc
1921 ccctccgcca cagtctgtgc ccccctctcc ccctccacct tcccccacca catacaactg
1981 tcccaggtcc ccgactccaa gagtctatgg gacaattaag cctgcgttca accagaaccc
2041 cgtcgccaag gtgccccccag ccaccaggtc tgacactgtg gccaccatga tgcgggaaaa
2101 ggggatgttc tacaggagag agctggaccg cttttccctg gactcagaag acgtctacag
2161 ccgcagcccc gccccacagg ctgccttccg caccaagcgg ggacagatgc ctgagaaccc
2221 gtactcggag gtgggaaaga tagccagcaa ggccgtctat gtcctgcca agccagccag
```

FIG. 2D-2

```
2281 gcggaagggc atgctggtaa agcagtccaa cgtggaggac agccctgaga agacgtgctc
2341 catcccatc ccaaccatca tagtcaagga accctccacc agcagcagcg gcaagagcag
2401 ccagggagc agcatggaga tcgacccca ggccactgag cccggccagc tgcggccaga
2461 tgacagcctc accgtcagca gcccttcgc ggcagccatc gctggggccg tgcgtgaccg
2521 ggagaagcgt ctggaagcca ggaggaattc cccggccttc ctctccaccg acctgggaga
2581 tgaggacgtg ggtctgggcc ctcctgctcc ccggatgcag ccctccaagt ttcccgagga
2641 gggtgggttt ggtgatgagg atgaaacgga acagccgcta ttgcctaccc ctggagcagc
2701 gcccagggag ctggagaacc acttcctagg tggtggtgag gctggtgctc aggtggaggc
2761 tgggggaccc ctgagttcca catccaaagc caagggggcct gaaagtggcc cagcagccgc
2821 cctcaagagc agcagcccag ccagccctga gaattacgtg cacccactca cggggcggct
2881 gctcgacccc agctccccgc tggccctggc actgtccgcc agggaccgag ccatgcagga
2941 gtcccagcaa ggacacaagg gggaggcccc caaggccgac cttaacaagc ctctctacat
3001 cgataccaaa atgcggccca gtgtgggagtc cggctttcca ccggtcacca gacagaacac
3061 cagggctccc ctgcgaaggc aggagacaga gaacaagtat gagacggacc tgaggcaagga
3121 ccggagggct gacgacaaga agaacatgct gatcaacatc gtggacactg cccagcagaa
3181 gtcagccggc ctactgatgg tgcacacagt ggacattccc gtggcggggc caccctggga
3241 ggaagaggag gacagagagg atggggatac aaagccagac cactcaccct ccacagtgcc
3301 agaaggcgtt cccaaaaccg aaggtgcttt acagatctcc gctgcccggg agcccgccgc
3361 tgcgcccggc aggaccatcg tggcagcggg ctccgtggaa gaggcggtaa ttctgccatt
3421 ccgcatcccc cctccccctc tggcgtccgt ggacttggat gaggactttc ttttcacaga
```

FIG. 2D-3

```
3481 accattgcct cctcccctgg aattcgccaa tagttttgat atccccgacg accgggcagc
3541 ttcagttccc gctctggctg acctggtcaa gcagaagaaa agtgacaccc ctcagccccc
3601 ttcgttgaac tccagccaac cagccaactc tacagacagt aagaagccag ccggcatctc
3661 aaactgtctg ccctcctcat tcctgccacc ccccgaaagc tttgacgcgg tcaccgactc
3721 ggggatcgag gaggtggaca gccggagtag cagcgaccac cacctggaga caaccagcac
3781 catctccaca gtgtccagca tctccacgct gtcttcagag ggcgggggaga gcatggatac
3841 ttgcacagtc tatgcagacg ggcaagcctt tgtggttgac aagccccccag tacctccaaa
3901 gccaaaaatg aagcccatcg ttcacaaaag caacgcactt taccaagaca cgctcccaga
3961 agaggacaca gatggctttg tgatcccccc acctgcaccc ccgccccccgc ccggcagtgc
4021 ccaggccggt gtggcgaagg tcatccagcc aaggacctcc aagttgtggg gtgatgttac
4081 agaggtcaaa agcccaattc tctcaggccc aaaggcaaat gtcattagtg agttaaactc
4141 cattctgcag cagatgaaca gggggaaatc ggtcaagcct ggggaagggc tggagctgcc
4201 cgtgggagcc aagtcagcta acctcgctcc aagaagcccg gaggtcatga gcactgtctc
4261 aggtacacgg agcacgacgg tcaccttcac tgtccgcccc ggaacctccc agcccatcac
4321 cctacagagc cggccccctg actatgaaag caggacctca ggaccgagac gtgccccaag
4381 ccctgtggtt tcgccaacgg aattgagcaa agagattctg cccacccctc cgtctgctgc
4441 agcagcctct ccctccccca cactctcaga tgtctttagc cttccaagcc agtccccctgc
4501 aggggacctc tttggcttga acccagcagg acgcagcagg tcaccatctc cttcaatatt
4561 gcaacagcca atctcaaata agccttttac aactaagcct gtccacctgt ggacgaaacc
4621 agatgtggca gactggctgg aaagtctgaa cttgggtgaa cacaaggaga cgttcatgga
```

FIG. 2D-4

```
4681 caatgagatt gacggcagcc atctgccaaa cctccagaag gaagacttga tagatcttgg
4741 ggtgactcga gttgggcaca ggatgaacat agaaagggct ttgaaacagc tgctggacag
4801 ataagggtgg ctgtcctggg ccttcacaga ctcctttctt ataagtagag atgggcttat
4861 gttgaaatgt gtggtggcca agcagaggct gacagcacca atctctggtc tgtgggtttg
4921 tctctgggta cccaaagaag tgctgtggat gtcccttggt gtggccaaac cagggcctca
4981 aaacccagct ctctgtgcag atttcttggg tggcttctgt ccaaagggac agtaagggggg
5041 tgtctgtcct ggtctcagag gctgccatag atggcatgtg ggacccctctc ctcctgtggc
5101 agataccagt gtactccaga tacctcccga ggcctccctc ctctgccttc tgggcaactc
5161 tcacccctca gccccggaac acagcctttc ttcaggcctt tggcctctgg gtgggttcgt
5221 tggctcggtc cctcccctgt gcttggctct cagtgagcag ccccaggtgg agccctgcct
5281 gagccagacc ttagagaaac actgtcccac ccacctctag tctgactctt gtgacacgca
5341 tctgaacttc tttggcccca aggtccttgg cagtggtggc tgtgtggaca ggatcgtgcc
5401 cagcttgcat aacttgcttt ctttacttct gcaaactgat tagcataatt ccagggtggc
5461 caaactgaag tcacacggag ttagtcaaag cacaaagtca cgatcctgag gaggagggag
5521 gcctggccat agagccagcc aatgtgtggc tgcccaagca cacaacctcc tccagggcaa
5581 ggacaggtgt ctgcccagag cacctggcca gaagggagag gtgggggctc ttgaaatttta
5641 ctctaattgg atgtctgcag agctccacag aggtcagaac aggtgagcat cttcttttac
5701 aaggacaaga agacccaagg ttggggggaa gcacggttgc catagagaag ctgttgtcct
5761 cccctggtgg cgggtagaat gtcacgccct tcactgcatc aggccacatg tgagcatcct
5821 gataaggcag cacttcggct ctcagaggcc aagaaaggtg gaactcaggg gccttctcc
```

FIG. 2D-5

```
5881 cttctctcag atagcggggga tcccaccggc ttgtcttaac atctggatct cagagaagcc
5941 tgtggtcagg acctggtgga gaacagaaag tggggcacag gctgtggcga agtttagatg
6001 tgtgcggaaa ggaaggtggg gagcatcctg gaccaagcct cacctcaggc ccggccacct
6061 gagcctaccg ttcttgcccc aagatctctc cctctaatac ctctaacatc ccagggaggt
6121 cgatcggctg cacacagcca ccacgcagcc agctaggctg ccgcaacccc ctctcccccа
6181 cactcagagg agactagaat gtcaggtgat ggtttttatca tttcagaact agcccagccc
6241 atctctaatt ataaaacatg ttttttcctt ttttttccctc cacttatgct cacagccgtg
6301 tgtctgacga ggcctggaac agcactacaa tgaataccat aaagtttagc aatttaaaaa
6361 aaaactttct ttactgcaag tttaaatacg tgtacagata gttcataagc acaatatctt
6421 aagaaaaaag tggccggtct accgggcagc ccctgtgcca cctagtgctg gaaagctaaa
6481 ggaaaaggggg ttgtggttta acaatactgt ttccttgggga caattataga agcaaccttt
6541 tcaaaattac ccttccctgg ccacttcctc agaacctgca acacggagtc tttcttcagc
6601 ttttgccaca acttcatcac tgaagggcaa ttatttcgtc aaaaggatta gaatccacga
6661 ctccacagtt tctccaacaa caacagacag acatgcatat tcgtgcatgc gtgcggacac
6721 acacacacac acacacacac acacacacac acgcacacac cctctcatgc atgcaccaca
6781 acccttgata tggagaccag ggctttgaac cttattctgt cccccaaacc ctttgcaggt
6841 caacccccaac ttctgaatcg tccaatgagc tggtaccagg tgaccgttga gcactttgct
6901 ctgggttctg tcattcccca cagcgaggcg aatgaatcag taatgctcca gagatgttga
6961 gtccagacac tgcgggtagc cactgaagag ttcagagagt gtgaaaatgt ggcccgccac
7021 taatttgtct tccaatgggt ggttcttcga agtccctctca gtgtgccaca gtactaggtc
```

FIG. 2D-6

```
7081 cctttcctgt gcctctggtt ggcattaacc tctgtgctca agtgccaagg cagatgctag
7141 gcaggagaga cctggtggct tgggggcccg ctgaacgtag gcagcaggtg acaggctgcg
7201 gactgccttc tatagacttc aagccacagg catccaggca gaagcaagct gactgccagc
7261 cccgggagct gataggccag ctttgctttg cagtgtctgg tcacagaaac acagctggac
7321 ctatcagaac tgacccctaa tcatccaaac acacatttgc ctcgatagtg ggtctctgtc
7381 tctgacagtt gatttttgcg cagaaagcaa tagtcccctca ttgaagttgc agcccactgt
7441 ccccccctcc catttggaac atgtcaggac caggccctac cccaggaccc tcctcagtgg
7501 tgaatcagat gttaaactgc tcataaggac catgttacca ctattccagg ttttaagatc
7561 aacttttgtt atatactgta atttaataac tgcataaatt tacatgccta gatttctgta
7621 atattgtgta tacaaaacca aaatctctca agctgtaaat tgtgtatacc cgccaggatc
7681 cccatttttgg ggcatccatg cgcatttttag atttaattca tacaattaaa aactcaatgt
7741 gactgctgac gtgctgaaac tttacataag aattatttgt gatactttag ttaataaaat
7801 ggtgaggttt ttttttcct gagttattga acaagcaaac attacccagc tgacctggca
7861 atgacttttt gtgtgtgtgc gggtaggggg ccacaaatat tgatttttt ttcctcccat
7921 tagcaatttt gggttttaaa tactaacatg gttcacacta tagtttgtgt gtaacaacac
7981 atgtcacatt atctttgttg ctgttaattt ctgtgctttt tttagactt tatttattta
8041 gaaggagaga gagaaagaga gagggagaga gagagggaga gagagcgaga gagaaagttc
8101 ctgtactttg cactttccc cccaaatcct tcaatctctt gtatggcaac caaaattaag
8161 taaaaaaaaa tatatataaa aa
```

FIG. 2D-7

SHANK 3 nucleotide sequence

```
   1 ttctgcggcc gcgctaaacc gtgccgccgt cgccgccgcc gctgcgcctg cggagccccc
  61 ggagccgctg tcccccgcgc tggccccggc cccggcccca tacggccccc tcccgcagta
 121 gcgcggtcgg cgggactctg gcgggggggtc agggggggcc agggcgccgc gcggagtccc
 181 cgtgcgctcc tctctccgcc gggaacagtc cgggcccccgg cgctagcacc gggatggacg
 241 gccccggggc cagcgccgtg gtcgtgcgcg tcggcatccc ggacctgcaa caaacgaagt
 301 gcctgcgtct ggatccaacc gcgcccgtgt gggccgccaa gcagcgtgtg ctctgcgccc
 361 tcaaccacag ccttcaggac gcgctcaact acgggctatt ccagcctccc tcccggggtc
 421 gcgccggcaa gttcctggat gaagagcggc tcttacagga ctacccgcct aacctggaca
 481 cgcccctgcc ctatctggag tttcgataca agcggagagt ttatgcccag aacctcatag
 541 atgacaagca gtttgcaaag ctgcacacaa aggcaaacct gaagaagttc atggactatg
 601 tccagctaca cagcacagac aaggtggccc gcctgctgga caaggggctg gaccccaatt
 661 tccatgaccc tgactcagga gagtgccctc tgagccttgc agcacagttg gacaacgcca
 721 ctgacctcct gaaggttctt cgcaatggcg gtgctcatct ggacttccga acccgagatg
 781 ggctaaccgc tgtccactgc gccacccgac agcggaatgc gggagcattg acgaccctgc
 841 tggacctggg ggcttcacct gactacaagg acagccgcgg cctgacgccc ctgtaccata
 901 gtgccctagg gggcggggat gccctctgct gtgagctgct tctccatgat cacgcacagt
 961 tggggaccac tgacgagaat ggctggcagg agatccatca ggcctgtcgc tttgggcatg
1021 tacagcactt ggagcacctg ctgttctatg gggccaacat gggtgcccag aacgcctcgg
```

FIG. 2E-1

```
1081 gaaacacagc cttgcacatc tgtgccctct ataaccagga gagctgtgcc cgcgtcctgc
1141 tttttccgtgg tgccaacaag gacgtccgca attacaacag ccagacagcc ttccaggtgg
1201 ccattattgc agggaacttt gagcttgccg aggtaatcaa gacccacaaa gactcggatg
1261 tcgtaccatt cagggaaacc cccagctatg caaagcgacg acgtctggct ggcccgagtg
1321 gcttggcatc ccctcggccc ttacagcgct cagccagtga tatcaacctg aagggtgacc
1381 agcccgcagc ttctcccggg cccactctcc gaagcctccc tcaccaactg ctgctccaga
1441 ggcttcagga ggagaaagac cgggacaggg atggtgagca ggagaacgac atcagcggtc
1501 cctcagcagg cagggggcgc cacagcaaga tcagccccag cgggcccggc ggatccggcc
1561 ccgcgcccgg ccccggcccg gcgtctcccg cgccccccgc gccgccgccc cggggcccga
1621 agcggaaact ttacagtgcc gtccccggcc gcaagttcat cgctgtgaag gcgcacagcc
1681 cgcagggcga gggcgagatc ccgctgcacc gcggcgaggc cgtgaaggtg ctcagcattg
1741 gggagggcgg ttctgggag ggaaccgtga agggccgtac aggctggttc ccagctgact
1801 gtgtggagga agtgcagatg cgacagtatg acacacggca tgaaactcga gaggaccgga
1861 cgaagcgtct ttccgccac tacactgtgg gttcctatga cagcctcact tcacacagtg
1921 attatgtcat tgatgataag gtggctatcc tgcaaaaacg ggaccatgag ggttttggct
1981 ttgttctccg gggagccaaa gcagagaccc ccattgagga gtttacaccc acacctgcct
2041 tccctgcgct ccagtacctt gagtctgtag atgtggaagg tgtggcctgg aaggctgggc
2101 ttcgcactgg ggacttcctc attgaggtaa acggagtgaa cgtcgtgaag gttggacaca
2161 agcaagtggt gggtctcatc cgtcagggtg gcaaccgtct ggtcatgaag gttgtgtctg
2221 ttaccaggaa gccagaggag gatagtgctc ggcgcagagc cccaccacct cccaagaggg
```

FIG. 2E-2

```
2281 cccccagcac cacgctgacc ctgcggtcca agtccatgac ggctgagctc gaggaactcg
2341 cttccattcg gagaaggaaa ggggagaagt tggatgagat cctggcggtt gctgcggaac
2401 caacgctgag gccagacatt gcagacgctg attccagggc agccactgtc aagcagcggc
2461 ccaccagccg gaggattacc cctgccgaga tcagctcatt gtttgagcga cagggcctcc
2521 cgggcccaga gaagctgccg ggctctctgc ggaaggggat tccacggacc aaatctgtag
2581 gggaggatga gaagctggca tccctactgg aagggcgttt cccacgcagc acatcaatgc
2641 aagacacagt gcgtgaaggc cgaggcattc cgcccccacc gcagaccgcc ccgccacccc
2701 cacccgcgcc ctactacttc gactccgggc cacccccccac cttctcacca ccgccaccac
2761 caccgggccg ggcctatgac actgtgcgct ccagcttcaa gccaggcctg gaggctcgtc
2821 tgggtgcagg ggcagctggc ctgtatgatt ctggcacacc tctgggcccg ctgccctacc
2881 ctgagcgcca gaagcgtgca cgctccatga tcatattgca ggactctgcg ccagaagtgg
2941 gcgatgtacc ccggcctgcg cctgcagcca caccgccctga gcgccccaag cgccggcctc
3001 ggccgtcagg ccctgatagt ccctatgcca acctgggcgc cttcagtgcc agcctctttg
3061 ctccgtcgaa accgcagcgc cgcaagagtc cgctggtgaa gcagcttcag gtggaggacg
3121 ctcaggagcg cgcggcgttg gccgtgggta gcccgggacc agtgggtgga agctttgcac
3181 gagaaccctc cccaacgcac cgcgggcccc gaccgggcgg ccttgactac agctctggag
3241 aaggcctggg gctcaccttt ggcgggccta gccctgggcc agtcaaggag cggcgcctgg
3301 aggagcgacg ccgttccact gtgttcctgt ctgtgggtgc catcgagggc aaccctccca
3361 gcgcggatct gccatcccta caaccctccc gctccattga tgagcgcctc ctggggacag
3421 gcgccaccac tggccgagat ttgctgctcc cctcccctgt ctctgctctg aagccattgg
```

FIG. 2E-3

```
3481 tcggtggtcc caaccttggg ccctcaagct ccaccttcat ccatcctctt actggcaaac
3541 ccttggatcc tagctcaccc ctagctcttg ctctggctgc ccgagagcgg gctctggcct
3601 cgcaaacacc ttcccggtcc cccacacccg tgcacagtcc tgatgctgac cgccctggac
3661 ccctctttgt ggatgtgcaa acccgagact ccgagagagg acccttggcc tccccagcct
3721 tctcccctcg gagtccagcc tggattccag tgcctgctcg cagagaggca gagaagccca
3781 ctcgggaaga gcggaagtca ccagaggaca agaaatccat gatcctcagc gtcttggaca
3841 cgtccttgca acggccagct ggcctcattg ttgtgcatgc caccagcaat ggacaggagc
3901 ccaacaggct ggggctgaa gaggagcgcc cgggtactcc ggagctggcc ccaaccccca
3961 tgcaggcagc agctgtggca gagcccatgc caagcccacg agcccaaccc cctggcaaca
4021 tcccagcaga tcccgggcca agccaaggca actcagagga ggagccaaag ctggtattcg
4081 ctgtgaacct gccacctgct caactgtcct ccaacgatga ggagaccaga gaggagctgg
4141 cccgcattgg gctagtgcca ccccctgaag agtttgccaa tgggatccctg ctggccaccc
4201 caccccagg accgggcccc ttgcccacca cggtacccag cccggcctca gggaagccca
4261 gcagcgagct gccccctgcc ccggagtctg cagctgactc tggagtagag gaggccgaca
4321 ctcgaagctc cagtgacccc cacctgggaga ccacaagcac catttccaca gtgtccagca
4381 tgtccacccct gagctcggag agtggagaac tcactgacac ccacacctcc tttgccgatg
4441 gacacacttt tctactcgag aagccaccag tgcctcccaa gcccaaactc aagtccccgc
4501 tgggggaaggg gccggtgacc ttcagggggcc cgctgctgaa gcaatcctcg gacagtgagc
4561 tcatggccca gcagcaccat gccacctcta ctgggttgac ttctgctgct gggcctgccc
4621 gccctcgcta cctcttccag agaaggtcca agctgtgggg ggacccccgtg gagagtcggg
```

FIG. 2E-4

```
4681 ggctcccctgg gcctgaggat gacaaaccaa ctgtgatcag tgagctcagc tcccgtctgc
4741 agcagctgaa taaagacact cgctccttgg gggaggaacc agttggtggc ctgggtagcc
4801 tgctggaccc tgctaagaag tcgcccattg cagcagctcg ctgcgcggtg gtcccgagtg
4861 ccggctggct cttcagcagc ctcggtgagc tgagcaccat ctcagcgcag cgcagccccg
4921 ggggcccggg cggagggccc tcctactcgg tgcgggcccag cggccggtac cccgtggcga
4981 gacgagcccc gagcccagtg aaacccgcat cgctggagcg ggtggagggg ctgggggcgg
5041 gcgtgggagg cgcgggcggg cccttcgggcc tcacgcctcc caccatcctc aagtcgtcca
5101 gcctctccat cccgcacgaa cccaaggaag tgcgcttcgt ggtgcgaagt gcgagtgcgc
5161 gcagccgctc cccctcacca tctccgctgc cctcgccttc tcctggctct ggccccagtg
5221 ccggcccgcg tcggccattt caacagaagc ccctgcagct ttggagcaag ttcgatgtgg
5281 gcgactggct ggagagcatc cacttaggcg agcaccgaga ccgcttcgag gaccatgaga
5341 tcgaaggcgc acacctgcct gcgctcacca aggaagactt cgtggagctg ggagtcacac
5401 gcgttggcca ccgcatgaac atcgagcgtg cgctcaggca gctggatggc agctgacgcc
5461 cctctccctc tcctgttcct gctgcgccct gccggcaggg cccccaccc tactccaggc
5521 cgcaggctcg gctcgccccc taccacgggcg cccgggccag gaatgttgca tgaatcgtcc
5581 tgtttgctgt tgcttggaga cttgccctgt acattgctta gtgcccctccc ctgccgctga
5641 acccccaccca gcacacagta agggcgcgga ccagggggggc tgggtggaag ggggttgggg
5701 cagggtgctc tggcctgacc acctccctcca cagctccctgg tggccattct tccagagggg
5761 gaacctagtc cagcatgcga ggtcaggaca cgccttggtg actcgggggg agggggagga
5821 cattgggggtt ctcgataggg gccaaggagc cccctgtttt acatatttta atccactcta
```

FIG. 2E-5

```
5881 tatttggaaa gagaaaagga acaaatatct ctgtccgtaa cagttcccgc cctcttcccc
5941 tcaagtcctc tcgctggtcc cgccacagct acccagtctt ccatctccgg cccctcactg
6001 ccacccccata tagggcaggg gacactccag ctggcctggg gttagccagg gtcctggcag
6061 cccacccctgg ggacccccgc tcagcccctt tccctcgctg agctatagta tgcccccaccc
6121 acccttttagg tgctgctcag ggggacgggt ggcaggcatt gcctgctggg cactagcagg
6181 gccaggtggc ctgggagatt attgcccctgg ggctgggccc cggtaaccca accccagcca
6241 tcatcttcac agggtctctc ccaaaggagg ggtctaacct ttccccactt cttgggcaac
6301 tacagcagag aagcctccct gcctcgcgcc ccaaagactc cccaattcct gccctgtgtg
6361 tgtgcaccac atgtgtgtgt gcacgcctgc gtgcttgtga aaattgggtg tggctgagcg
6421 catgggtgcc ctgtatgtgc ttgattgtgg agtggtcccc aggggctgtt ctggatgggt
6481 gggaggttga ggaagcttgc acaggggtgc atgcatgggt gtgtgcctgt gaaagggccc
6541 tgtcttctcc aaagaaaggc tgtcctgctc ttgggtcctg ctgttttctc agcctgttct
6601 ccctgaacct cacccagctt aagcaggggt tcttggtgaa tcctttcagc tttgggaggc
6661 ctcaagggct cccgtgcagg cagcacccct ttgggcttct aagggaattg tggggaccac
6721 taaaatcagg ccacaacagc ccttggagag aggcaaagac tcctgagggt accctggccc
6781 cccttactgt gactcctcac aattcagcaa tgacctgtgg ggcgggggcc cttgggggcat
6841 tttaacata gggtttggag tctggactaa gctccatcca cgtcactcac aagtttctgt
6901 ttctatttct agctttttt aataaaatat atatatatat atataaaaa gacagaaaac
6961 aggtgttttc atggcccagg ggcttggcac gccggtctgt gcccacccgc cccgccccac
7021 cctggcccac cggccccatt ccttagacac agagtcacgc ccactaaccc tcttaccaac
```

FIG. 2E-6

```
7081 agagcaggtc acacacacag cagcggtcac tgtaacagac tgccacatac acagtctcac
7141 atttacctgt gggtttttgg ttctgttcag tttgggtttt taactttaca gggtcagttc
7201 cgcttcatcc ccctttttgta tggagttcca tctcgggggct ttcaaccccc tgctccagtc
7261 ctgaggcctc ctgaccctga cgttgtgata caccccacag agatctatgt ttcttatatt
7321 attattatta ataataatta ttataatatt atgtaataaa tttataagaa atgaaaaaaa
7381 aaaaaaaaaa a
```

FIG. 2E-7

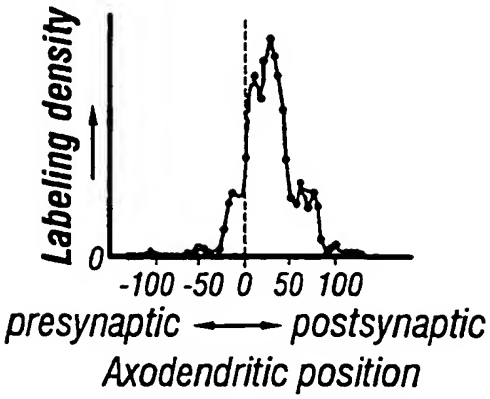

FIG. 3A

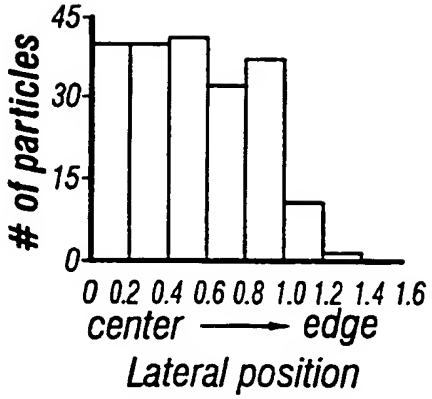

FIG. 3B

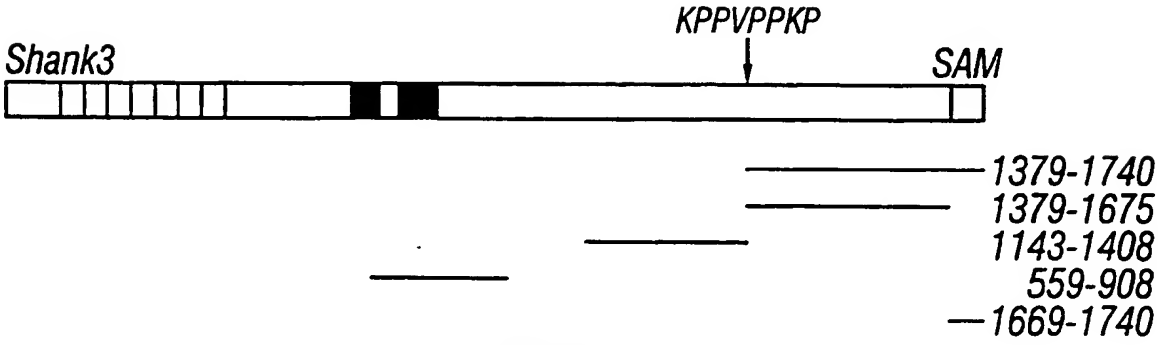

FIG. 4

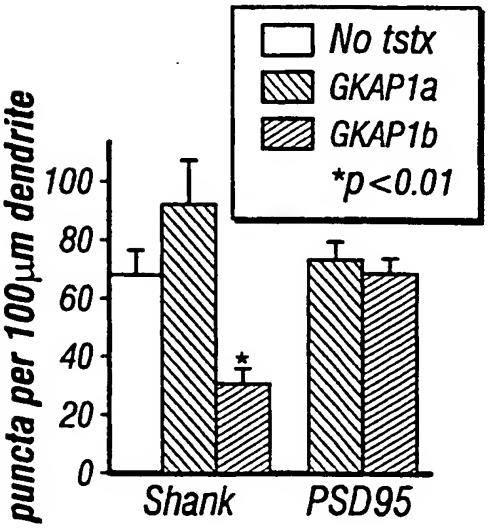

FIG. 5A

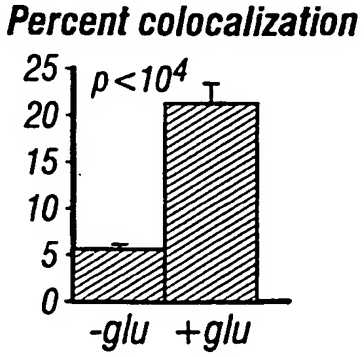

FIG. 5B

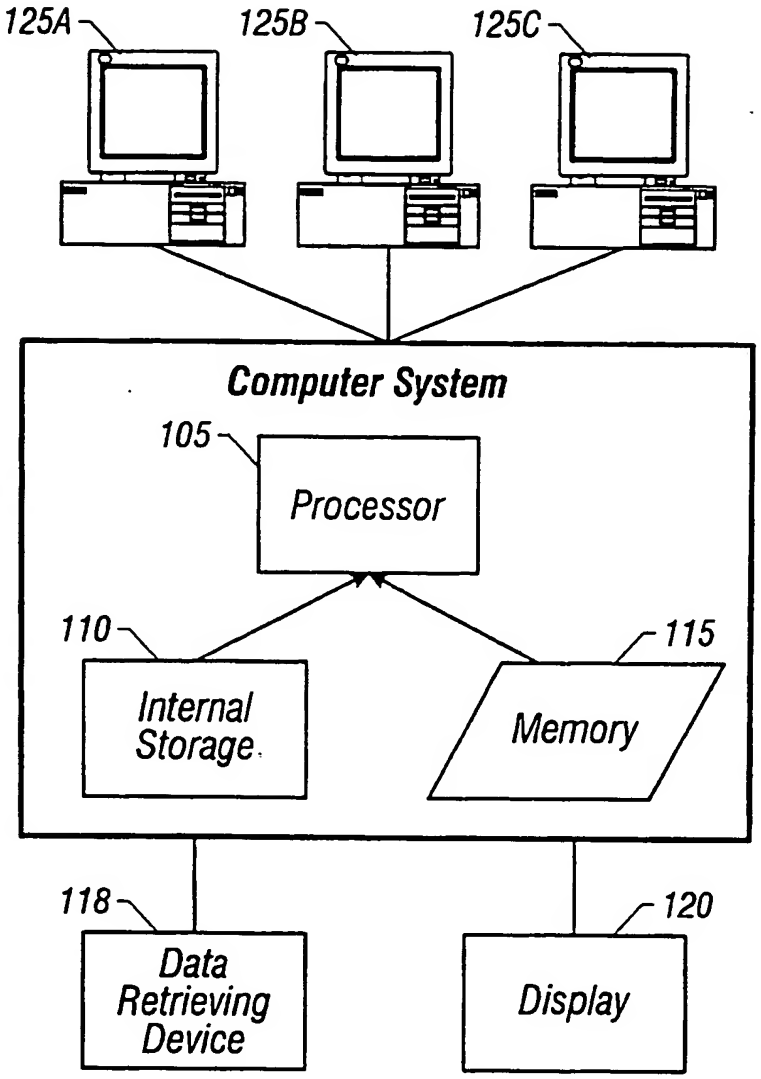
125A
125B
125C
Computer System
105
Processor
110
115
Internal Storage
Memory
118
120
Data Retrieving Device
Display

FIG. 6

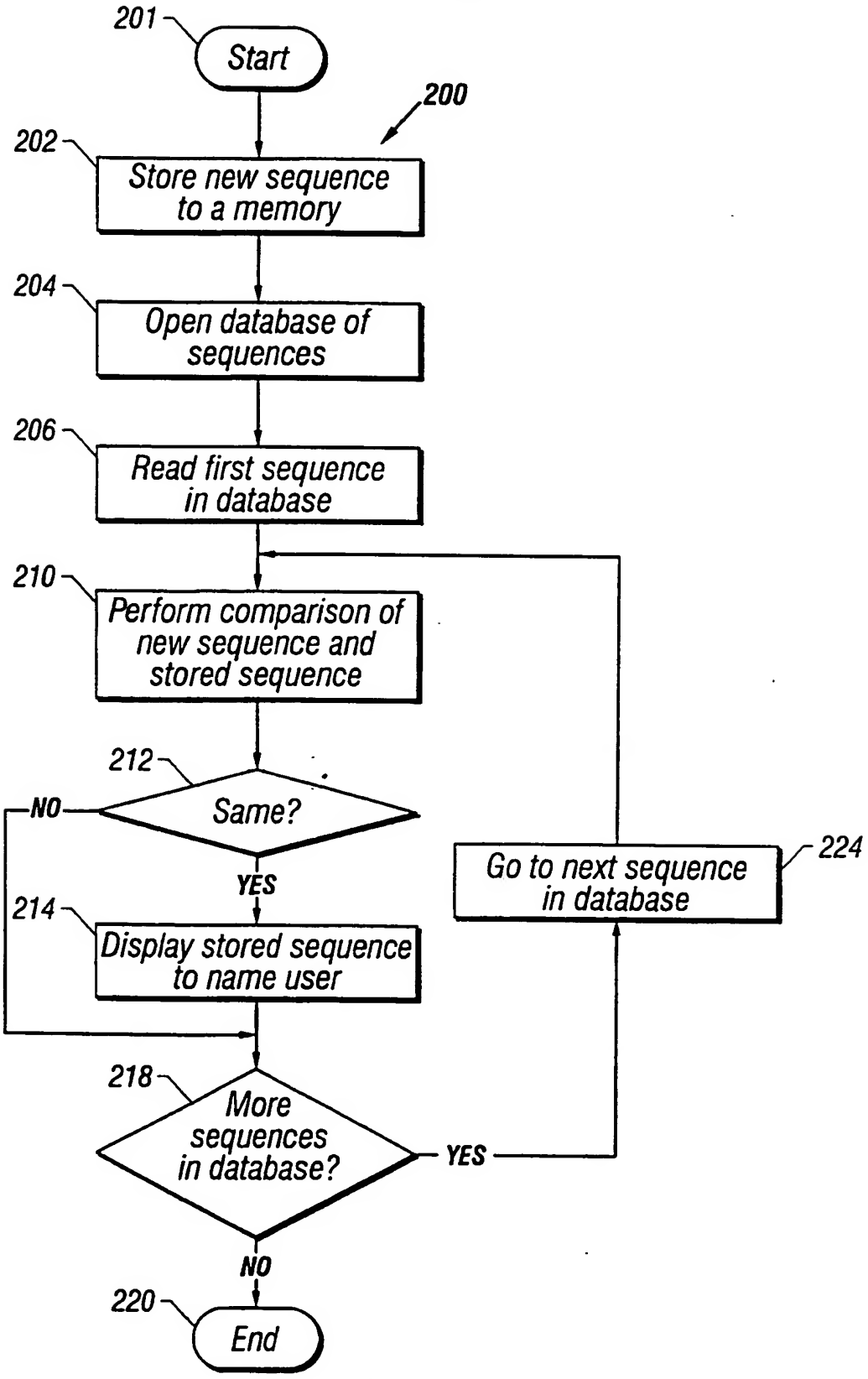
201
Start
200
202
Store new sequence to a memory
204
Open database of sequences
206
Read first sequence in database
210
Perform comparison of new sequence and stored sequence
212
Same?
NO
YES
214
Display stored sequence to name user
224
Go to next sequence in database
218
More sequences in database?
YES
NO
220
End

FIG. 7

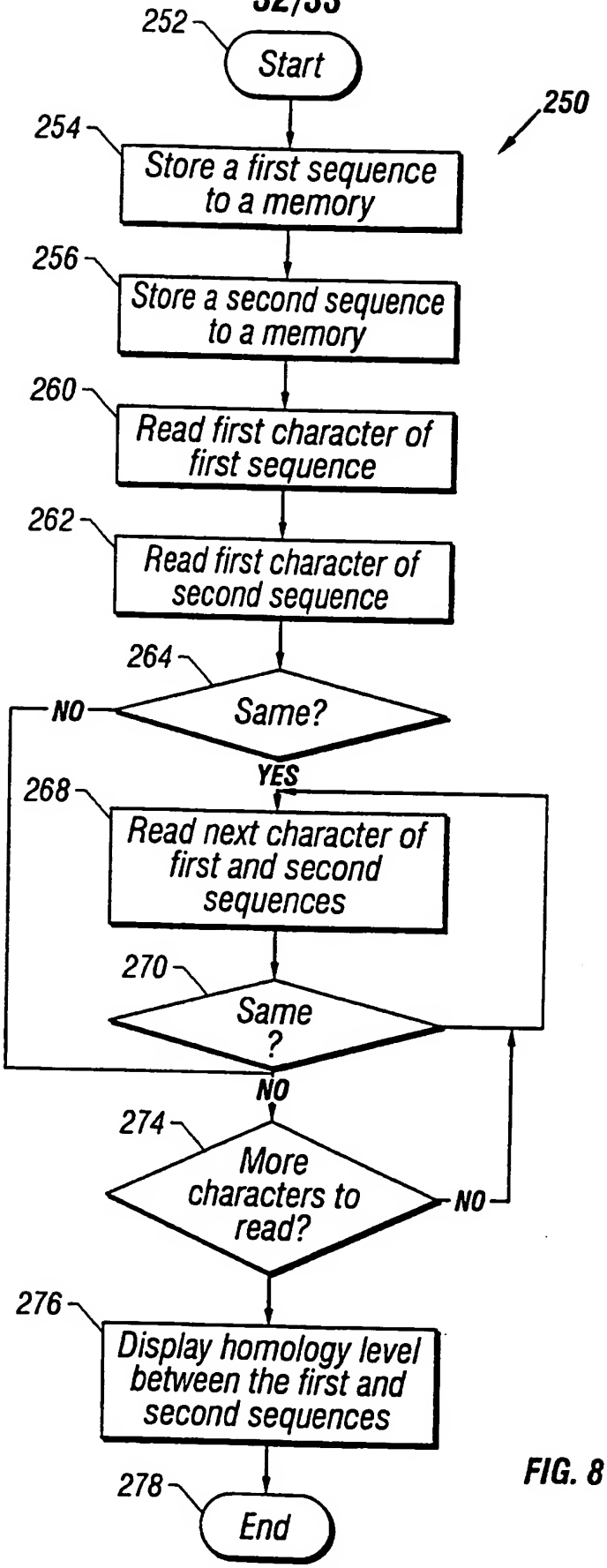
252
Start
250
254
Store a first sequence to a memory
256
Store a second sequence to a memory
260
Read first character of first sequence
262
Read first character of second sequence
264
Same?
NO
YES
268
Read next character of first and second sequences
270
Same ?
NO
274
More characters to read?
NO
276
Display homology level between the first and second sequences
278
End

FIG. 8

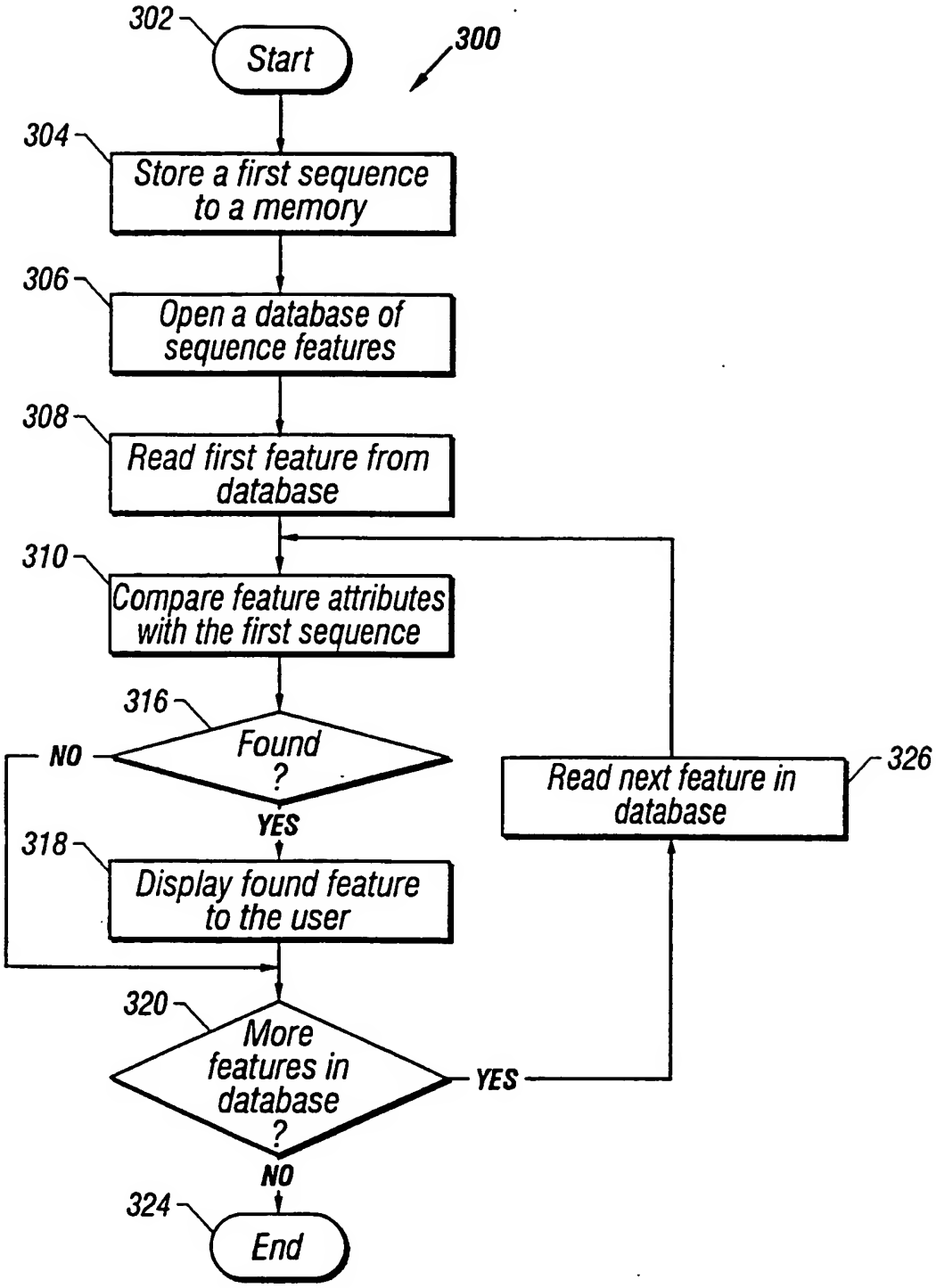
302
Start
300
304
Store a first sequence to a memory
306
Open a database of sequence features
308
Read first feature from database
310
Compare feature attributes with the first sequence
316
Found ?
NO
YES
318
Display found feature to the user
320
More features in database ?
YES
NO
324
End
326
Read next feature in database

FIG. 9